MC-HGNN: MARKOVIAN SEMANTIC TRANSPORT FOR HETEROGENEOUS GRAPH LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Heterogeneous graphs contain multiple node and relation types, and predictive information often depends on long-range semantic dependencies formed by sequences of relations. Existing heterogeneous Graph Neural Networks (GNNs) typically capture this structure through predefined meta-paths or by stacking local message-passing and attention layers. These approaches increase computational cost, can amplify hub-dominated propagation, and make deep models susceptible to over-smoothing. To address this, we propose **MC-HGNN**, a Markovian semantic transport framework. It interprets relation-aware graph attention as conditional typed transitions and couples them through learned, node-dependent relation selection. A hub-aware correction limits excessive transport toward highly influential nodes. Relation-dependent continuation lets different semantic channels learn different propagation ranges. The resulting all-hop transport implicitly aggregates variable-length heterogeneous paths without explicit meta-path enumeration. It also decouples transport depth from neural-network depth. Theoretically, we establish adaptive meta-path composition, an exponentially convergent finite transport approximation, distance-decaying robustness to localized perturbations, and quantitative resistance to over-smoothing. Experiments on seven heterogeneous graph benchmarks show that MC-HGNN consistently improves predictive performance while maintaining favorable memory and computational efficiency.

1 INTRODUCTION

Many real-world graphs are heterogeneous, containing multiple node and relation types whose interactions can determine the prediction task. For example, in an academic network, a paper's label may depend not only on its direct neighbors but also on higher-order typed patterns such as paper-author-paper or paper-subject-paper. We refer to such task-relevant information as *predictive structural signal*, i.e., relation-aware connectivity patterns that help distinguish node classes or preserve semantic similarity. In heterogeneous graphs, this signal is rarely confined to a single edge. Instead, it is often distributed across multi-step typed dependencies, commonly described as meta-paths (Li et al., 2024b). Efficiently capturing this long-range semantic structure therefore remains a central challenge.

Existing heterogeneous graph learning methods only partially address this problem. Relational message-passing models (Wang et al., 2019; Schlichtkrull et al., 2018) capture longer dependencies by stacking multiple local aggregation layers, but increasing depth can lead to over-smoothing and can progressively weaken discriminative node information. Heterogeneous graph transformers (Hu et al., 2020; Yang et al., 2023) improve local expressivity through relation-aware attention, yet long-range reasoning still largely depends on repeatedly applying expensive attention blocks (Yu et al., 2022). Meta-path-based methods can explicitly encode higher-order semantics, but they typically require predefined path templates and may miss useful relation sequences that are not specified in advance. These limitations become more severe on large graphs, where deep attention, auxiliary structural encodings, and repeated neighborhood aggregation increase both memory and computational cost. In addition, heterogeneous graphs often contain highly connected hubs that can attract disproportionate attention and dominate repeated propagation.

Our key observation is that row-normalized graph attention can be interpreted as a probabilistic transition over neighboring nodes. This opens a natural route to long-range reasoning. Rather than

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

repeatedly recomputing full attention blocks, the learned local transitions can be reused to propagate information over multiple steps. In heterogeneous graphs, however, meaningful trajectories typically involve changing relation types. Therefore, long-range transport must adapt both the relation selected at each node and the distance over which information is allowed to travel. This motivates a Markovian formulation in which relation-aware attention defines local typed transitions, while learned relation selection and continuation jointly govern the evolution of the transport process.

Based on this perspective, we propose **MC-HGNN**, which constructs a heterogeneous Markov process by factorizing each transition into node-dependent relation selection and relation-conditioned neighbor selection. Hub-aware reweighting limits excessive transport toward high-degre nodes, while relation-dependent continuation determines whether a trajectory should extend after each semantic relation. The resulting all-hop operator aggregates adaptive, variable-length typed trajectories without explicit meta-path enumeration. Because the learned transitions are reused for sparse multi-step transport, MC-HGNN separates transport depth from attention depth. This enables long-range heterogeneous propagation without repeatedly recomputing full attention blocks. This formulation also yields theoretically grounded and interpretable guarantees for adaptive meta-path composition, efficient finite approximation, robustness to localized perturbations, and resistance to over-smoothing.

Our main contributions are: (i) We introduce a Markovian formulation of heterogeneous graph attention that couples relation-specific typed transitions into an adaptive multi-relation transport process. (ii) We propose hub-aware reweighting, node-dependent relation selection, and relation-specific continuation to capture variable-length heterogeneous dependencies without predefined meta-path templates or deep attention stacking. (iii) We establish theoretical results on implicit adaptive meta-path composition, efficient approximation of all-hop transport, robustness to localized perturbations, and quantitative resistance to over-smoothing. (iv) We evaluate MC-HGNN on seven heterogeneous graph benchmarks and show consistent predictive improvements while maintaining favorable memory and runtime efficiency.

2 Related Work

Heterogeneous Graph Neural Networks (GNNs). Early heterogeneous GNNs used predefined meta-paths. Heterogeneous Graph Attention Network (HAN) (Wang et al., 2019) applies semantic-level attention over meta-path views, while Metapath Aggregated Graph Neural Network (MAGNN) (Fu et al., 2020) improves intra-path aggregation. Relational Graph Convolutional Network (R-GCN) (Schlichtkrull et al., 2018) learns relation-specific propagation, but remains one-hop per layer. Thus, these methods capture multi-hop semantics mainly through depth, which increases the risk of over-smoothing. MC-MC-HGNN addresses this depth–over-smoothing trade-off using a discounted resolvent operator that aggregates multiple propagation lengths within a single relation-aware layer.

Graph transformers for heterogeneous data. Heterogeneous Graph Transformer (HGT) (Hu et al., 2020) and Simple Heterogeneous Graph Neural Network (Simple-HGN) (Yang et al., 2023) often require positional or structural encodings for long-range topology. Long-range Meta-path Search and Propagation Strategy (LMSPS) (Li et al., 2024a) captures long semantic paths through path enumeration. MC-HGNN addresses this computational bottleneck using relation-wise all-hop transport operator directly from learned attention without deep stacking or positional encodings.

3 Preliminaries

Heterogeneous Graphs and Meta-Paths. Let $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \tau, \phi)$ be a heterogeneous graph, where $\tau : \mathcal{V} \to \mathcal{T}$ maps nodes to node types and $\phi : \mathcal{E} \to \mathcal{R}$ maps edges to relation types. The topological structure is encoded by relation-specific adjacency matrices $\mathbf{A}_r \in \{0,1\}^{|\mathcal{V}|\times|\mathcal{V}|}$. In such environments, predictive structural signals often depend on composite sequences of relations, or **meta-paths**. A meta-path specifies a typed relation schema $\mathcal{P} : \tau_1 \xrightarrow{r_1} \tau_2 \xrightarrow{r_2} \cdots \xrightarrow{r_k} \tau_{k+1}$, where $\tau_i \in \mathcal{T}$ and $r_i \in \mathcal{R}$. A concrete node sequence $v_1 \xrightarrow{r_1} v_2 \cdots \xrightarrow{r_k} v_{k+1}$ that follows this schema is referred to as a meta-path instance. For example, $\mathrm{Author} \to \mathrm{Paper} \to \mathrm{Venue}$ describes publication-mediated author–venue semantics.

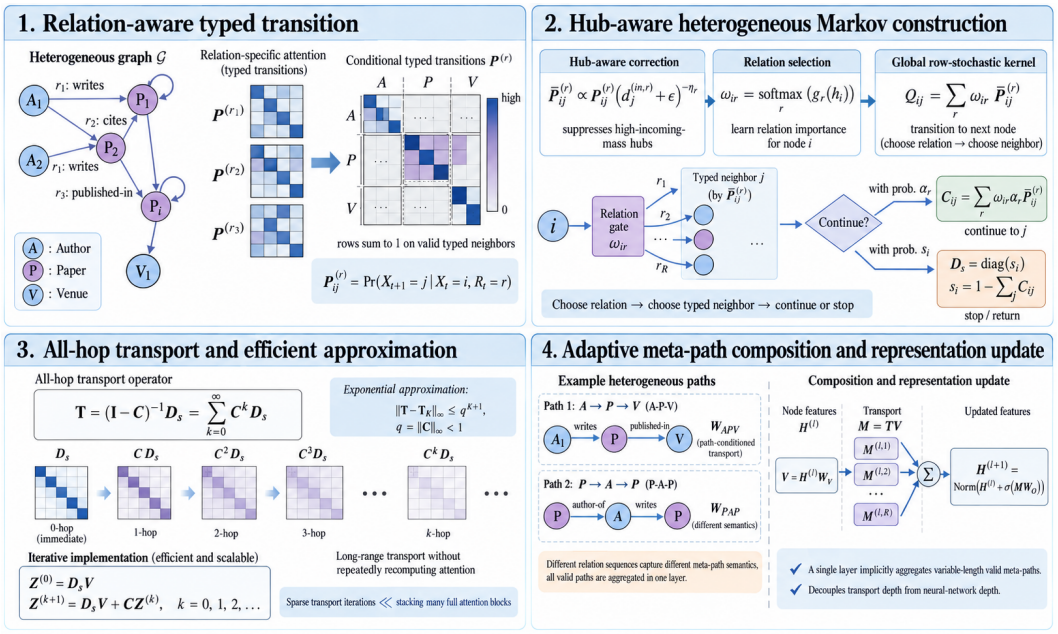


Figure 1: Overview of MC-HGNN. Relation-aware attention is interpreted as a row-stochastic Markov transition kernel. A resolvent operator aggregates all-hop typed transport.

Relation-Aware Attention. Modern architectures such as HGT (Hu et al., 2020) capture typed dependencies through **relation-aware attention**. For each relation r and layer ℓ, let $h_i^{(\ell)} \in \mathbb{R}^d$ denote the hidden representation of node i. We compute typed query $q_i^{(\ell,r)}$, key $k_j^{(\ell,r)}$, and value $u_j^{(\ell,r)}$ vectors as $q_i^{(\ell,r)} = W_{Q,\tau(i)}^{(\ell,r)} h_i^{(\ell)}$, $k_j^{(\ell,r)} = W_{K,\tau(j)}^{(\ell,r)} h_j^{(\ell)}$, $u_j^{(\ell,r)} = W_{V,\tau(j)}^{(\ell,r)} h_j^{(\ell)}$, where $W_{Q,\tau(i)}^{(\ell,r)}$, $W_{K,\tau(j)}^{(\ell,r)}$, and $W_{V,\tau(j)}^{(\ell,r)}$ are learnable type- and relation-specific projection matrices. Let $\mathcal{N}_r(i) = \{j : (i,j) \in \mathcal{E}, \phi(i,j) = r\}$ denote the r-typed neighbors of node i. For each $j \in \mathcal{N}_r(i)$, the attention score is computed as

$$e_{ij}^{(\ell,r)} = \frac{\langle q_i^{(\ell,r)}, k_j^{(\ell,r)} \rangle}{\sqrt{d}}, \tag{1}$$

where $\langle \cdot, \cdot \rangle$ denotes the standard inner product and d is the hidden feature dimension.

Markov Transition Kernels. For each relation $r \in \mathcal{R}$, let $\mathbf{A}_r$ denote the adjacency matrix associated with relation r. Let $\mathbf{D}_r$ be its diagonal degree matrix with $(\mathbf{D}_r)_{ii} = \sum_j (\mathbf{A}_r)_{ij}$. A natural row-stochastic transition operator for relation r is defined as $\mathbf{P}_r = \mathbf{D}_r^{-1} \mathbf{A}_r$. Since each row of $\mathbf{P}_r$ sums to one, $\mathbf{P}_r$ can be interpreted as a Markov transition kernel that describes one-step movement of probability mass along edges of type r. For a sequence of relations $(r_1, r_2, \ldots, r_k)$, the corresponding multi-step typed transition operator is given by $\mathbf{P}^{(k)} = \mathbf{P}_{r_1} \mathbf{P}_{r_2} \cdots \mathbf{P}_{r_k}$. This represents transport along the prescribed typed relation sequence.

4 Methodology

Problem Setup. Given a heterogeneous graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \tau, \phi)$ defined in Section 3, our goal is to learn node representations that capture both local typed interactions and long-range semantic dependencies. MC-HGNN interprets relation-aware attention as conditional transition probabilities over typed graph edges and couples these transitions into a heterogeneous Markov process. The resulting transition operator is then used to propagate information over adaptive semantic paths within each network layer. Figure 1 summarizes the overall pipeline. Given node representations $\mathbf{H}^{(\ell)}$, MC-HGNN constructs relation-specific attention kernels, applies hub-aware correction, and learns which relation to follow from each node. Relation-dependent probabilities then control how far

semantic information propagates. A resolvent aggregates transport over paths of different lengths. Finally, the representation is combined with the input representation to obtain $\mathbf{H}^{(\ell+1)}$.

4.1 Relation-Aware Attention as Typed Transition

For each relation $r \in \mathcal{R}$, relation-specific query and key representations are constructed as defined above, and the corresponding attention scores $e_{ij}^{(\ell,r)}$ are computed over valid typed neighbors. We normalize these scores using a row-wise softmax over the valid r-neighbors of node i. This produces the transition weights $P_{ij}^{(\ell,r)}$, with $P_{ij}^{(\ell,r)} = 0$ for $j \notin \mathcal{N}_r(i)$. For every node with at least one outgoing edge of relation r, the resulting weights satisfy $\sum_j P_{ij}^{(\ell,r)} = 1$. We interpret this normalized attention as the conditional transition probability $P_{ij}^{(\ell,r)} = \Pr(X_{t+1} = j \mid X_t = i, R_t = r)$, where X_t denotes the node occupied by the transport process at step t and R_t denotes the selected relation. We write $\mathbf{P}^{(\ell,r)} \in \mathbb{R}^{|\mathcal{V}|\times|\mathcal{V}|}$ for the sparse matrix with entries $(\mathbf{P}^{(\ell,r)})_{ij} = P_{ij}^{(\ell,r)}$. Thus, relation-aware attention determines how semantic information moves among valid neighbors once a relation has been selected. Each $\mathbf{P}^{(\ell,r)}$ is represented as a sparse $|\mathcal{V}| \times |\mathcal{V}|$ operator whose nonzero entries occur only on edges belonging to relation r. Rows corresponding to nodes without an outgoing r-edge remain zero. This common node-space representation allows different relation transitions to be combined while preserving their typed edge supports.

Example. Consider three papers (A, B, C) and two authors (Alice, Bob) connected through the "author-of" relation. For clarity, we display only the Paper-to-Author block of the full sparse $|\mathcal{V}| \times |\mathcal{V}|$ transition operator: $\mathbf{P}^{(\text{author-of})} = \begin{bmatrix} 1.0 & 0.0 \\ 0.6 & 0.4 \\ 0.0 & 1.0 \end{bmatrix}$. Conditioned on selecting "author-of", Paper A transitions to Alice, Paper B transitions to Alice or Bob with probabilities 0.6 and 0.4, and Paper C transitions to Bob. A subsequent transition can select another compatible relation, e.g., $\text{Paper} \xrightarrow{\text{author-of}} \text{Author} \xrightarrow{\text{writes}} \text{Paper}$. Hence, long-range semantic transport naturally consists of sequences of potentially different relation types.

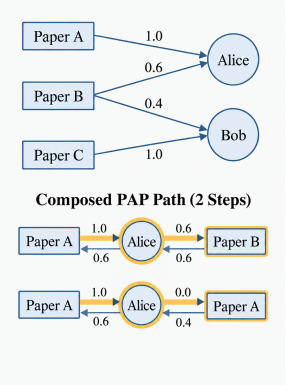


Figure 2: Example of typed semantic transport.

Hub-Aware Transition Reweighting. Heterogeneous graphs often contain high in-degree nodes that receive a disproportionate amount of attention mass. Repeated transport through such nodes can dominate information propagation and suppress less-connected regions of the graph. We therefore correct the relation-specific transition probabilities before constructing the global Markov process.

For relation r, let $d_j^{(\text{in},r)} = \sum_i P_{ij}^{(\ell,r)}$ denote the incoming transition mass received by node j. We define

$$\bar{P}_{ij}^{(\ell,r)} = \frac{P_{ij}^{(\ell,r)} \left(d_j^{(\text{in},r)} + \epsilon\right)^{-\eta_r}}{\sum_{k\in\mathcal{N}_r(i)} P_{ik}^{(\ell,r)} \left(d_k^{(\text{in},r)} + \epsilon\right)^{-\eta_r}}, \tag{2}$$

where $\epsilon > 0$ ensures numerical stability and $\eta_r = \sigma(\xi_r) \in (0, 1)$ is learned jointly with the model, with $\sigma(x) = 1/(1 + e^{-x})$ the sigmoid function and $\xi_r \in \mathbb{R}$ a learnable scalar. When η_r is close to zero, the original attention distribution is approximately preserved. Increasing η_r suppresses destinations receiving large incoming transition mass. The row normalization preserves $\sum_j \bar{P}_{ij}^{(\ell,r)} = 1$ whenever relation r is available from node i.

4.2 Heterogeneous Markov Transition

A node in a heterogeneous graph may participate in several semantic relations. We therefore learn not only which neighbor should receive information, but also which relation should govern the next transition. Let $\mathcal{R}(i) = \{r : \mathcal{N}_r(i) \neq \emptyset\}$ denote the relations available from node i. For nodes with $\mathcal{R}(i) = \varnothing$, we use the self-transition $Q_{ii}^{(\ell)} = 1$. Equivalently, in the continuation process we set

$\mathbf{C}^{(\ell)}_{i,:} = \mathbf{0}$ and $s^{(\ell)}_i = 1$. We assign each available relation a node-dependent probability $\omega^{(\ell)}_{ir} = \frac{\exp(g_r(\mathbf{h}^{(\ell)}_i))}{\sum_{s\in\mathcal{R}(i)}\exp(g_s(\mathbf{h}^{(\ell)}_i))}$, where $g_r(\cdot)$ is a learnable relation-scoring function. Therefore, $\omega^{(\ell)}_{ir} \geq 0$ and $\sum_{r\in\mathcal{R}(i)}\omega^{(\ell)}_{ir} = 1$. The relation-selection probability and the conditional edge transition together define

$$Q^{(\ell)}_{ij} = \sum_{r\in\mathcal{R}(i)} \omega^{(\ell)}_{ir}\bar{P}^{(\ell,r)}_{ij}. \tag{3}$$

Equivalently, $\omega^{(\ell)}_{ir} = \Pr(R_t = r \mid X_t = i)$ and $\bar{P}^{(\ell,r)}_{ij} = \Pr(X_{t+1} = j \mid X_t = i, R_t = r)$. Marginalizing over relation choices gives $Q^{(\ell)}_{ij} = \Pr(X_{t+1} = j \mid X_t = i)$. Since each $\bar{\mathbf{P}}^{(\ell,r)}$ is conditionally row-normalized, we have $\sum_j Q^{(\ell)}_{ij} = \sum_r \omega^{(\ell)}_{ir} = 1$. Hence, $\mathbf{Q}^{(\ell)}$ defines a row-stochastic heterogeneous Markov transition kernel. The powers of $\mathbf{Q}^{(\ell)}$ encode valid multi-step heterogeneous trajectories. In particular, an entry of $(\mathbf{Q}^{(\ell)})^k$ aggregates paths of the form $i = v_0 \xrightarrow{r_1} v_1 \xrightarrow{r_2} \cdots \xrightarrow{r_k} v_k = j$, where each (v_{t-1}, v_t) belongs to the edge set of relation r_t. Thus, the transport process can change relation type at every step while remaining constrained by the heterogeneous graph structure.

4.3 Relation-Dependent Continuation

Not all relations should propagate information over the same effective distance. We therefore associate each relation r with a learnable continuation probability α_r. We parameterize $\alpha_r = \underline{\alpha} + (\overline{\alpha} - \underline{\alpha})\sigma(a_r)$, where $0 \leq \underline{\alpha} < \overline{\alpha} < 1$ are fixed continuation bounds. We denote the largest learned continuation probability by $\alpha_\star = \max_r \alpha_r$, so that $\alpha_\star \leq \overline{\alpha} < 1$. At node i, the process first selects a relation r according to $\omega^{(\ell)}_{ir}$. With probability α_r, it follows an edge sampled from $\bar{\mathbf{P}}^{(\ell,r)}$. Otherwise, the transport terminates at the current node. Marginalizing over relation choices gives the continuation kernel

$$C^{(\ell)}_{ij} = \sum_{r\in\mathcal{R}(i)} \omega^{(\ell)}_{ir}\alpha_r\bar{P}^{(\ell,r)}_{ij}. \tag{4}$$

Its row sum is $c^{(\ell)}_i = \sum_j C^{(\ell)}_{ij} = \sum_r \omega^{(\ell)}_{ir}\alpha_r < 1$. The remaining mass $s^{(\ell)}_i = 1 - c^{(\ell)}_i$ is therefore the probability that transport terminates at node i. We collect these probabilities in $\mathbf{D}^{(\ell)}_s = \mathrm{diag}(s^{(\ell)}_1, \ldots, s^{(\ell)}_{|\mathcal{V}|})$.

The process can equivalently be represented as an absorbing Markov chain with one terminal copy of each graph node, $\mathcal{M}^{(\ell)} = \begin{bmatrix} \mathbf{C}^{(\ell)} & \mathbf{D}^{(\ell)}_s \\ \mathbf{0} & \mathbf{I} \end{bmatrix}$. The first block contains transient graph states, while the second contains their absorbing terminal copies. Thus, termination at node j is explicitly represented by absorption into the terminal copy of j.

4.4 All-Hop Semantic Transport

MC-HGNN aggregates transport over trajectories of different lengths rather than selecting a fixed propagation radius. The resulting operator is

$$\mathbf{T}^{(\ell)} = \sum_{k=0}^{\infty}\left(\mathbf{C}^{(\ell)}\right)^k \mathbf{D}^{(\ell)}_s = \left(\mathbf{I} - \mathbf{C}^{(\ell)}\right)^{-1}\mathbf{D}^{(\ell)}_s. \tag{5}$$

Since every row of $\mathbf{C}^{(\ell)}$ sums to at most $\alpha_\star \leq \overline{\alpha} < 1$, we have $\|\mathbf{C}^{(\ell)}\|_\infty \leq \alpha_\star < 1$. Therefore, the Neumann series converges and Eq. 5 is well defined.

Moreover, since $\mathbf{D}^{(\ell)}_s\mathbf{1} = (\mathbf{I} - \mathbf{C}^{(\ell)})\mathbf{1}$, we obtain $\mathbf{T}^{(\ell)}\mathbf{1} = \mathbf{1}$. Hence, $\mathbf{T}^{(\ell)}$ is itself row-stochastic, and $T^{(\ell)}_{ij}$ can be interpreted as the total probability that a semantic walk beginning at node i eventually terminates at node j.

For a path $\mathcal{P} : v_0 \xrightarrow{r_1} v_1 \xrightarrow{r_2} \cdots \xrightarrow{r_k} v_k$, its contribution to the transport operator is

$$w(\mathcal{P}) = \left[\prod_{t=1}^{k} \omega^{(\ell)}_{v_{t-1}r_t} \alpha_{r_t} \bar{P}^{(\ell,r_t)}_{v_{t-1}v_t}\right] s^{(\ell)}_{v_k}. \tag{6}$$

Therefore, the contribution of a semantic path depends jointly on edge-level attention, relation preference, relation-dependent continuation, and the probability of terminating at its final node.

Representation Update and Efficient Approximation. Before transport, node representations are projected into a common value space as $\mathbf{V}^{(\ell)} = \mathbf{H}^{(\ell)}\mathbf{W}^{(\ell)}_V$. The ideal transported representation is $\mathbf{M}^{(\ell)} = \mathbf{T}^{(\ell)}\mathbf{V}^{(\ell)}$.

We do not explicitly construct the matrix inverse in Eq. 5. Instead, the transport is approximated using $K_{\max}$ sparse iterations:

$$\mathbf{Z}^{(0)} = \mathbf{D}^{(\ell)}_s \mathbf{V}^{(\ell)}, \qquad \mathbf{Z}^{(k+1)} = \mathbf{D}^{(\ell)}_s \mathbf{V}^{(\ell)} + \mathbf{C}^{(\ell)}\mathbf{Z}^{(k)}. \tag{7}$$

After $K_{\max}$ iterations, $\mathbf{Z}^{(K_{\max})} = \sum_{k=0}^{K_{\max}} (\mathbf{C}^{(\ell)})^k \mathbf{D}^{(\ell)}_s \mathbf{V}^{(\ell)}$, which converges to $\mathbf{M}^{(\ell)}$ as $K_{\max}$ increases. The node representation is then updated as

$$\mathbf{H}^{(\ell+1)} = \mathrm{Norm}\left(\mathbf{H}^{(\ell)} + \sigma\left(\mathbf{Z}^{(K_{\max})}\mathbf{W}^{(\ell)}_O\right)\right). \tag{8}$$

Within each MC-HGNN layer, the relation-specific attention kernels, relation-selection probabilities, and continuation kernel are computed once from $\mathbf{H}^{(\ell)}$ and remain fixed throughout the $K_{\max}$ transport iterations. The next MC-HGNN layer recomputes these quantities from the updated node representations. This separates the depth of semantic transport from the number of nonlinear attention layers: a shallow stack of MC-HGNN layers can perform long-range propagation through repeated sparse Markov transitions. Finally, $\mathbf{C}^{(\ell)}$ does not need to be materialized as a dense matrix. Its action on a feature matrix can be computed directly as $\mathbf{C}^{(\ell)}\mathbf{Z} = \sum_r \mathbf{\Omega}^{(\ell)}_r \alpha_r \bar{\mathbf{P}}^{(\ell,r)}\mathbf{Z}$, where $\mathbf{\Omega}^{(\ell)}_r = \mathrm{diag}(\omega^{(\ell)}_{1r}, \ldots, \omega^{(\ell)}_{|\mathcal{V}|r})$. Constructing relation-aware edge scores, hub corrections, and relation selection requires $O(|\mathcal{E}|d)$ sparse computation, after which the $K_{\max}$ transport iterations require $O(K_{\max}|\mathcal{E}|d)$. Thus, the overall block complexity is $O((K_{\max}+1)|\mathcal{E}|d)$ up to relation-gating costs, with $O(|\mathcal{E}|)$ sparse transition storage and no dense all-pairs matrix.

4.5 Theoretical Properties

Proposition 1 (Implicit Adaptive Meta-Path Composition) *For any nodes i and j, the all-hop transport operator satisfies $T^{(\ell)}_{ij} = \sum_{k=0}^{\infty} \sum_{\mathcal{P}\in\mathcal{P}_k(i,j)} w^{(\ell)}(\mathcal{P})$, where $\mathcal{P}_k(i,j)$ denotes the valid typed path instances of length k from i to j and $w^{(\ell)}(\mathcal{P})$ is defined in Eq. 6. Hence, MC-HGNN implicitly aggregates variable-length heterogeneous path instances corresponding to different meta-path schema without requiring the schema to be predefined.*

Theorem 1 (Exponential Approximation of All-Hop Transport) *Let $\mathbf{T}^{(\ell)}$ be the all-hop transport operator and let $\mathbf{T}^{(\ell)}_K = \sum_{k=0}^{K} (\mathbf{C}^{(\ell)})^k \mathbf{D}^{(\ell)}_s$ be its K-step approximation. Define $q_\ell = \|\mathbf{C}^{(\ell)}\|_\infty = \max_i \sum_j C^{(\ell)}_{ij} < 1$. Then $\|\mathbf{T}^{(\ell)} - \mathbf{T}^{(\ell)}_K\|_\infty \le q_\ell^{K+1}$. Consequently, for any value representation $\mathbf{V}$, $\|(\mathbf{T}^{(\ell)} - \mathbf{T}^{(\ell)}_K)\mathbf{V}\|_{\infty,2} \le q_\ell^{K+1}\|\mathbf{V}\|_{\infty,2}$. Hence, the finite sparse transport used by MC-HGNN converges exponentially to the infinite-horizon operator as K increases.*

Theorem 1 shows that the infinite-horizon transport does not need to be computed exactly in practice. The contribution of paths beyond K steps is bounded by q_ℓ^{K+1}, so a small number of sparse transport iterations can closely approximate the full resolvent whenever $q_\ell < 1$.

Theorem 2 (Quantitative Resistance to Oversmoothing) *Let $\mathcal{C} = \{\mathbf{1}\mathbf{a}^\top : \mathbf{a} \in \mathbb{R}^d\}$ denote the consensus subspace and define the node-wise contrast $\Delta(\mathbf{V}) = \inf_{\mathbf{Y}\in\mathcal{C}} \|\mathbf{V} - \mathbf{Y}\|_{\infty,2}$., where $\|\mathbf{X}\|_{\infty,2} = \max_i \|\mathbf{x}_i\|_2$. For an MC-HGNN transport block, let $q_\ell = \|\mathbf{C}^{(\ell)}\|_\infty < 1$. Then, $\Delta\big(\mathbf{T}^{(\ell)}\mathbf{V}\big) \ge \frac{1-q_\ell}{1+q_\ell}\Delta(\mathbf{V}) \ge \frac{1-\alpha_\star}{1+\alpha_\star}\Delta(\mathbf{V})$. Consequently, after L finite transport blocks, $\Delta(\mathbf{V}^{(L)}) \ge \left[\prod_{\ell=1}^{L} \frac{1-q_\ell}{1+q_\ell}\right] \Delta(\mathbf{V}^{(0)}) > 0$ whenever $\Delta(\mathbf{V}^{(0)}) > 0$. Thus, the linear Markov transport component preserves a nonzero fraction of node-wise contrast after any finite number of transport blocks.*

Table 1: Main results on seven heterogeneous node-classification benchmarks (Macro-F1 / Micro-F1). Results are averaged over 10 runs on ACM, DBLP, IMDB, Freebase, and AMiner, and 5 runs on Yelp and OGBN-MAG. **Standard deviations are reported in Appendix F**

Method	ACM		DBLP		IMDB		Freebase		OGBN-MAG		Yelp		AMiner	
	Ma	Mi	Ma	Mi	Ma	Mi	Ma	Mi	Ma	Mi	Ma	Mi	Ma	Mi
R-GCN	90.80	91.70	91.20	92.10	57.60	59.40	59.80	61.10	48.70	50.20	63.20	64.80	85.60	86.90
HAN	91.90	92.80	92.60	93.40	58.90	60.80	61.40	62.60	49.80	51.50	65.10	66.50	87.40	88.50
MAGNN	92.70	93.50	93.40	94.10	60.40	62.20	62.60	63.80	50.60	52.10	66.00	67.30	88.10	89.20
HGT	93.40	94.10	94.20	95.00	61.70	63.30	63.50	64.60	51.90	53.60	67.30	68.50	89.00	90.10
Simple-HGN	93.10	93.90	93.80	94.70	61.20	62.90	63.00	64.10	51.40	53.00	66.80	68.10	**90.90**	**91.80**
SeHGNN	93.50	94.20	94.30	95.00	62.00	63.50	63.80	64.90	**52.10**	53.80	67.50	68.80	89.20	90.30
LMSPS	93.60	94.30	94.40	95.10	62.10	63.70	64.00	**66.10**	51.80	**55.40**	67.80	**70.00**	89.50	90.50
MC-HGNN	**94.40**	**95.00**	**95.10**	**95.80**	**64.00**	**65.50**	**66.20**	65.40	**53.60**	55.20	**69.40**	**70.60**	89.80	89.70

Theorem 2 shows that MC-HGNN preserves a nonzero fraction of node-wise contrast after each transport block, rather than allowing representation to collapse.

Theorem 3 (Distance-Decaying Robustness to Localized Noise) *Let $\mathbf{V} = \mathbf{V}^* + \mathbf{E}$, where $\mathbf{E}$ is supported on a corrupted node set $\mathcal{S}$ and $\|\mathbf{e}_j\|_2 \le \epsilon$ for $j \in \mathcal{S}$ ($\mathbf{e}_j$ denotes the jth row of $\mathbf{E}$). Let $d(i, \mathcal{S})$ denote the shortest valid heterogeneous transport distance from node i to $\mathcal{S}$, and let $\alpha_\star = \max_r \alpha_r < 1$. For a fixed MC-HGNN transport operator, let $\mathbf{M} = \mathbf{TV}$ and $\mathbf{M}^* = \mathbf{TV}^*$, and let $\mathbf{m}_i$ and $\mathbf{m}_i^*$ denote their ith rows. Then $\|\mathbf{m}_i - \mathbf{m}_i^*\|_2 \le \epsilon\, \alpha_\star^{d(i,\mathcal{S})}$. Hence, the influence of localized perturbations decays exponentially with heterogeneous transport distance.*

5 Experiments

Datasets. We evaluate MC-HGNN on seven heterogeneous node-classification benchmarks: academic/citation networks (ACM, DBLP, AMiner, and OGBN-MAG), the Freebase knowledge graph, the IMDB movie network, and the Yelp user–business network. OGBN-MAG is the largest benchmark, with approximately 1.9M nodes and 21M edges. Dataset statistics and train/validation/test splits are provided in Appendix E.1.

Baselines. We compare MC-HGNN with representative heterogeneous graph methods spanning relation-specific message passing, meta-path modeling, heterogeneous attention, and scalable propagation. The baselines include R-GCN (Schlichtkrull et al., 2018), HAN (Wang et al., 2019), MAGNN (Fu et al., 2020), HGT (Hu et al., 2020), Simple-HGN (Yang et al., 2023), and LMSPS (Li et al., 2024a).

Evaluation Protocol. All methods use the same dataset splits, optimizer family, early-stopping criterion, and comparable model-capacity budget. We report mean Macro-F1 and Micro-F1 over 10 runs on ACM, DBLP, IMDB, Freebase, and AMiner, and over 5 runs on Yelp and OGBN-MAG due to their larger computational cost. Hyperparameters are selected using validation performance only. The learned continuation parameters α_r and hub-correction parameters η_r are optimized jointly with the model. Full training and search details are provided in Appendix E.

5.1 Main Results

Table 1 summarizes performance across the seven benchmarks. MC-HGNN achieves the best Macro-F1 on six of seven datasets, with particularly strong gains on Freebase, Yelp, and OGBN-MAG. On Freebase, MC-HGNN improves Macro-F1 from 64.0 to 66.2, while on Yelp it improves from 67.8 to 69.4. On OGBN-MAG, MC-HGNN obtains 53.6 Macro-F1, exceeding the strongest baseline by 1.5 points in Macro-F1. The gains are smaller on ACM and DBLP, where local neighborhoods are already highly informative. These results suggest that the proposed transport is most useful when prediction requires integrating information across multi-step heterogeneous dependencies.

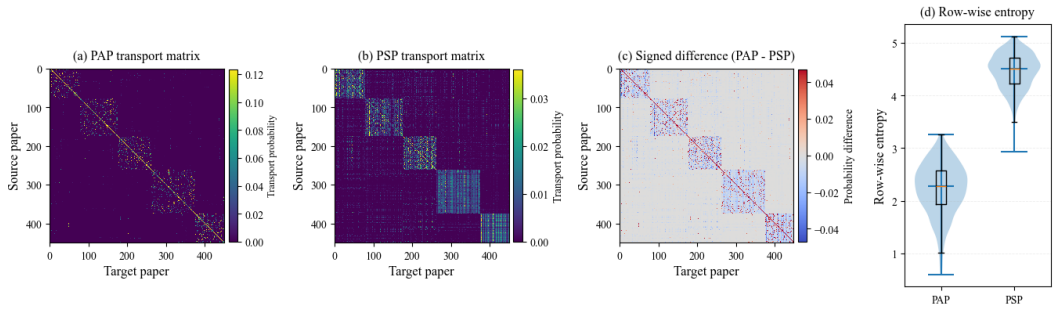


Figure 3: Relation-specific meta-path transport patterns on ACM: (a) PAP transport matrix, (b) PSP transport matrix, (c) signed difference map, and (d) row-wise entropy distribution.

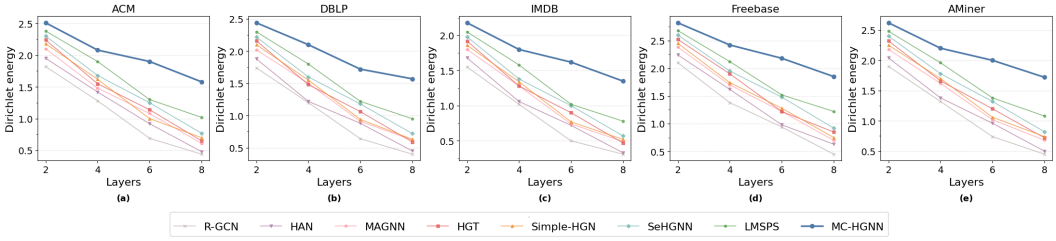


Figure 4: Dirichlet energy versus depth on heterogeneous benchmarks. Higher retained energy indicates slower loss of representation variation under repeated propagation.

Adaptive Heterogeneous Transport Analysis. We examine whether MC-HGNN learns distinct transport patterns for different semantic relation sequences. Let $\mathbf{C}_r = \mathbf{\Omega}_r \alpha_r \bar{\mathbf{P}}_r$ denote the contribution of relation r to the continuation kernel, with $\mathbf{C} = \sum_r \mathbf{C}_r$. For the common ACM meta-paths PAP (paper–author–paper) and PSP (paper–subject–paper), we extract $\mathbf{W}_{\mathrm{PAP}} = \mathbf{C}_{PA}\mathbf{C}_{AP}\mathbf{D}_s$, $\mathbf{W}_{\mathrm{PSP}} = \mathbf{C}_{PS}\mathbf{C}_{SP}\mathbf{D}_s$. For visualization, nonzero rows are normalized to unit mass and the same semantic ordering of papers is used in all panels; this ordering affects only the display. Figure 3 shows that PAP and PSP induce distinct transport structures, while the signed difference map highlights source–target pairs emphasized differently by the two channels.

We quantify transport selectivity using the row-wise entropy $H_i = -\sum_j \widehat{W}_{ij} \log \widehat{W}_{ij}$. PAP has a lower mean entropy than PSP (4.03 vs. 5.27), indicating a more concentrated transport distribution. Together, the structural and entropy differences show that MC-HGNN assigns distinct transport behavior to different heterogeneous relation sequences.

Table 2: Robustness under hub-centered perturbation (Macro-F1).

	IMDB			*Yelp*		
Method	Clean	Hub_5	Hub_{10}	Clean	Hub_5	Hub_{10}
HGT	61.7	57.9 (-3.8)	55.2 (-6.5)	67.3	63.1 (-4.2)	60.4 (-6.9)
Simple-HGN	61.2	57.5 (-3.7)	54.8 (-6.4)	66.8	62.7 (-4.1)	60.0 (-6.8)
LMSPS	62.1	58.8 (-3.3)	56.4 (-5.7)	67.8	64.5 (-3.3)	62.0 (-5.8)
MC-HGNN	**64.0**	**61.3 (-2.7)**	**59.4 (-4.6)**	**69.4**	**67.0 (-2.4)**	**65.1 (-4.3)**

5.1.1 Over-Smoothing and Hub Domination Analysis

Over-smoothing. Figure 4 shows how Dirichlet energy (Cai & Wang, 2020) changes with depth on five heterogeneous benchmarks. Across all datasets, the competing methods exhibit a rapid decline in energy as the number of layers increases, indicating progressively stronger smoothing of node representations. In contrast, MC-HGNN consistently preserves higher Dirichlet energy, suggesting that it better retains non-smooth and discriminative components as propagation deepens. Although Dirichlet energy differs from the node-wise contrast measure in Theorem 2, the result provides complementary empirical evidence that MC-HGNN avoids rapid over-smoothing.

Sensitivity to hub-dominated structure. For each graph, we identify the top-$k\%$ highest-degree nodes and rewire 50% of their incident edges. Each rewired edge preserves its relation type and source–destination node-type constraint, while its endpoint is replaced by another valid destination of the corresponding type. This perturbs hub-centered connectivity without introducing schema-invalid edges. As a result, it provides a controlled test of whether a model relies excessively on hub-dominated connectivity patterns. Table 2 reports results on IMDB and Yelp under two perturbation levels, Hub_5 and Hub_{10}. Although performance degrades for all methods as the perturbation strength increases, MC-HGNN consistently achieves the strongest Macro-F1 and exhibits smaller degradation from the clean graph. This supports the claim that hub-aware reweighting reduces over-reliance on dominant hubs and preserves useful relational structure under hub-centered corruption.

Localized feature corruption. We further evaluate Theorem 3 by perturbing the value representations of a randomly selected node set $\mathcal{S}$ while keeping the learned transport operator fixed. For every uncorrupted node, we measure the resulting representation change and group nodes by their shortest directed heterogeneous transport distance to $\mathcal{S}$. The perturbation decreases with distance and remains below the theoretical envelope $\epsilon\alpha_\star^{d(i,\mathcal{S})}$. Detailed results are provided in Appendix F.3.

Ablation study. Table 3 isolates the main components of MC-HGNN. Removing multi-hop transport causes the largest degradation, confirming the importance of long-range heterogeneous propagation. Uniform relation selection and shared continuation also reduce performance, showing that both node-dependent relation choice and relation-specific propagation ranges are useful. Hub correction provides consistent gains, while the matched-PPR control shows that the improvement is not explained solely by generic all-hop diffusion.

Table 3: Ablation study (Macro-F1).

Variant	DBLP	IMDB	Yelp	OGBN-MAG
MC-HGNN	**95.1**	**64.0**	**69.4**	**53.6**
w/o hub correction	94.5	63.2	68.2	52.7
Uniform relation selection	94.2	62.8	67.8	52.3
Shared continuation	94.5	63.2	68.3	52.8
One-hop attention	93.6	61.8	66.9	51.5
Matched PPR	94.6	63.3	68.5	52.9

Scalability at scale. Table 4 shows that the sparse transport formulation of MC-HGNN translates into practical efficiency on OGBN-MAG. MC-HGNN requires only 8.9 GB of peak GPU memory and 18.5 s per training epoch, corresponding to a 30.3% reduction in memory and a $1.87\times$ speedup relative to HGT. Under the reported configuration, this yields lower peak memory and training time than the attention-based baselines while retaining competitive predictive performance.

Table 4: Scalability and efficiency on OGBN-MAG.

Method	Peak Mem	Mem Red.	Time/Ep	Speedup
HGT	12.8 GB	–	34.6 s	$1.00\times$
Simple-HGN	10.7 GB	16.4%	28.9 s	$1.20\times$
LMSPS	11.3 GB	11.7%	30.1 s	$1.15\times$
MC-HGNN	**8.92 GB**	**30.3%**	**18.5 s**	**1.87**$\times$

Limitations and Future Work MC-HGNN extends its propagation range through sequential sparse transport, so increasing $K_{\max}$ introduces additional sparse matrix-feature multiplications that cannot be fully parallelized within a transport block. The current formulation assumes a static graph topology. Extending Markovian transport to continuously evolving heterogeneous graphs remains an interesting direction.

6 Conclusion

We introduced **MC-HGNN**, a Markovian semantic transport framework for heterogeneous graph learning. MC-HGNN interprets relation-aware attention as conditional typed transitions and couples them through node-dependent relation selection, hub-aware reweighting, and relation-dependent continuation. The resulting transport process aggregates variable-length heterogeneous trajectories while separating propagation range from nonlinear network depth. Our analysis characterizes implicit meta-path composition, finite-step approximation, localized perturbation decay, and preservation of node-wise contrast. Experiments across seven heterogeneous benchmarks demonstrate strong predictive performance together with favorable scalability.

Reproducibility Statement

Complete proofs of all theoretical results are provided in the appendix. Appendix E reports the dataset statistics, evaluation protocols, training configuration, model implementation, and hyperparameter settings. Additional ablations, transport diagnostics, robustness analyses, and finite-transport approximation results are also provided in the appendix.

AI Use Statement

Generative AI tools were used to assist with manuscript grammar check and paraphrasing.

References

Chen Cai and Yusu Wang. A note on over-smoothing for graph neural networks. *arXiv preprint arXiv:2006.13318*, 2020.

Xinyu Fu, Jiani Zhang, Ziqiao Meng, and Irwin King. Magnn: Metapath aggregated graph neural network for heterogeneous graph embedding. In *The Web Conference*, 2020.

Ziniu Hu, Yuxiao Dong, Kuansan Wang, and Yizhou Sun. Heterogeneous graph transformer. In *The Web Conference*, 2020.

Chao Li, Zijie Guo, Qiuting He, and Kun He. Long-range meta-path search on large-scale heterogeneous graphs. In *Advances in Neural Information Processing Systems*, 2024a.

Munan Li, Kai Liu, Hongbo Liu, Zheng Zhao, Tomas E Ward, and Xindong Wu. Heterogeneous meta-path graph learning for higher-order social recommendation. *ACM Transactions on Knowledge Discovery from Data*, 18(8):1–25, 2024b.

Michael Schlichtkrull, Thomas N. Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Welling. Modeling relational data with graph convolutional networks. In *European Semantic Web Conference*, 2018.

Xiao Wang, Houye Ji, Chuan Shi, Bai Wang, Yanfang Ye, Peng Cui, and Philip S. Yu. Heterogeneous graph attention network. In *The World Wide Web Conference*, 2019.

Xiaocheng Yang, Mingyu Yan, Shirui Zhang, Xiaochun Ye, Dongrui Li, Chuan Yang, James Fan, Hui Huang, and Zhong Hu. Simple and efficient heterogeneous graph neural network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023.

Le Yu, Leilei Sun, Bowen Du, Chuanren Liu, Weifeng Lv, and Hui Xiong. Heterogeneous graph representation learning with relation awareness. *IEEE Transactions on Knowledge and Data Engineering*, 35(6):5935–5947, 2022.

A Proof of Proposition 1

Recall that the heterogeneous continuation kernel at layer ℓ is

$$\mathbf{C}^{(\ell)} = \sum_{r\in\mathcal{R}} \mathbf{\Omega}_r^{(\ell)} \alpha_r \bar{\mathbf{P}}^{(\ell,r)}, \tag{9}$$

where $\mathbf{\Omega}_r^{(\ell)} = \mathrm{diag}(\omega_{1r}^{(\ell)}, \ldots, \omega_{|\mathcal{V}|r}^{(\ell)})$. Therefore, its (u, v)-th entry is

$$C_{uv}^{(\ell)} = \sum_{r\in\mathcal{R}} \omega_{ur}^{(\ell)} \alpha_r \bar{P}_{uv}^{(\ell,r)}. \tag{10}$$

We first consider transport over exactly k steps. By matrix multiplication,

$$\left[(\mathbf{C}^{(\ell)})^k\right]_{ij} = \sum_{v_1,\ldots,v_{k-1}} \prod_{t=1}^{k} C_{v_{t-1}v_t}^{(\ell)}, \tag{11}$$

where $v_0 = i$ and $v_k = j$. Substituting Eq. 10 into Eq. 11 gives

$$\left[(\mathbf{C}^{(\ell)})^k\right]_{ij} = \sum_{v_1,\ldots,v_{k-1}} \prod_{t=1}^{k} \left(\sum_{r_t \in \mathcal{R}} \omega^{(\ell)}_{v_{t-1} r_t} \alpha_{r_t} \bar{P}^{(\ell,r_t)}_{v_{t-1} v_t} \right). \tag{12}$$

Expanding the product over relation choices yields

$$\left[(\mathbf{C}^{(\ell)})^k\right]_{ij} = \sum_{\mathcal{P} \in \mathcal{P}_k(i,j)} \prod_{t=1}^{k} \omega^{(\ell)}_{v_{t-1} r_t} \alpha_{r_t} \bar{P}^{(\ell,r_t)}_{v_{t-1} v_t}, \tag{13}$$

where $\mathcal{P} = (v_0 = i, r_1, v_1, \ldots, r_k, v_k = j)$. By construction, $\bar{P}^{(\ell,r)}_{uv} = 0$ whenever $(u, v) \notin \mathcal{E}_r$. Consequently, every term corresponding to an invalid typed transition vanishes, and only edge-valid heterogeneous relation sequences contribute to Eq. 13.

The all-hop transport operator is

$$\mathbf{T}^{(\ell)} = \sum_{k=0}^{\infty} (\mathbf{C}^{(\ell)})^k \mathbf{D}^{(\ell)}_s. \tag{14}$$

Since $\mathbf{D}^{(\ell)}_s$ is diagonal, its multiplication assigns the terminal factor $s^{(\ell)}_j$ to every path ending at node j. Thus,

$$T^{(\ell)}_{ij} = \sum_{k=0}^{\infty} \sum_{\mathcal{P} \in \mathcal{P}_k(i,j)} \left[\prod_{t=1}^{k} \omega^{(\ell)}_{v_{t-1} r_t} \alpha_{r_t} \bar{P}^{(\ell,r_t)}_{v_{t-1} v_t} \right] s^{(\ell)}_j. \tag{15}$$

Defining

$$w^{(\ell)}(\mathcal{P}) = \left[\prod_{t=1}^{k} \omega^{(\ell)}_{v_{t-1} r_t} \alpha_{r_t} \bar{P}^{(\ell,r_t)}_{v_{t-1} v_t} \right] s^{(\ell)}_j, \tag{16}$$

Eq. 15 is exactly the path expansion stated in Proposition 1.

Therefore, each power $(\mathbf{C}^{(\ell)})^k$ aggregates all valid heterogeneous paths of length k, while the resolvent sums these contributions over all $k \geq 0$. Since the weights depend on the current node, selected relation, learned edge transition, and relation-dependent continuation probability, the contribution of each path is adaptive rather than fixed by a predefined meta-path template. Hence, a single MC-HGNN layer implicitly composes variable-length heterogeneous meta-path dependencies. This completes the proof.

B Proof of Theorem 3

Recall that the all-hop transport operator is

$$\mathbf{T}^{(\ell)} = \sum_{k=0}^{\infty} (\mathbf{C}^{(\ell)})^k \mathbf{D}^{(\ell)}_s, \tag{17}$$

where $\mathbf{C}^{(\ell)}$ is nonnegative and satisfies $\|\mathbf{C}^{(\ell)}\|_\infty \leq \alpha_\star < 1$.

Let $\mathbf{M} = \mathbf{T}^{(\ell)}\mathbf{V}$ and $\mathbf{M}^* = \mathbf{T}^{(\ell)}\mathbf{V}^*$, and denote their ith rows by $\mathbf{m}_i$ and $\mathbf{m}^*_i$, respectively. Since $\mathbf{V} = \mathbf{V}^* + \mathbf{E}$,

$$\mathbf{M} - \mathbf{M}^* = \mathbf{T}^{(\ell)}\mathbf{E}. \tag{18}$$

For node i, using the fact that $\mathbf{E}$ is supported only on $\mathcal{S}$ gives

$$\mathbf{m}_i - \mathbf{m}^*_i = \sum_{j \in \mathcal{S}} T^{(\ell)}_{ij} \mathbf{e}_j. \tag{19}$$

Applying the triangle inequality and $\|\mathbf{e}_j\|_2 \leq \epsilon$, we obtain

$$\|\mathbf{m}_i - \mathbf{m}^*_i\|_2 \leq \epsilon \sum_{j \in \mathcal{S}} T^{(\ell)}_{ij}. \tag{20}$$

We next bound the transport mass reaching the corrupted set. Let $d = d(i, \mathcal{S})$. By definition, there is no valid heterogeneous path from i to any node in $\mathcal{S}$ having length smaller than d. Since $(\mathbf{C}^{(\ell)})^k$ contains only valid length-k transport paths,

$$\left[(\mathbf{C}^{(\ell)})^k \mathbf{D}_s^{(\ell)}\right]_{ij} = 0, \qquad j \in \mathcal{S}, \;\; k < d. \tag{21}$$

Therefore,

$$\begin{aligned}\sum_{j\in\mathcal{S}} T_{ij}^{(\ell)} &= \sum_{k=d}^{\infty} \sum_{j\in\mathcal{S}} \left[(\mathbf{C}^{(\ell)})^k \mathbf{D}_s^{(\ell)}\right]_{ij} \\ &\le \sum_{k=d}^{\infty} \sum_{j} \left[(\mathbf{C}^{(\ell)})^k \mathbf{D}_s^{(\ell)}\right]_{ij}.\end{aligned} \tag{22}$$

Factoring out the first d transitions gives

$$\sum_{k=d}^{\infty} (\mathbf{C}^{(\ell)})^k \mathbf{D}_s^{(\ell)} = (\mathbf{C}^{(\ell)})^d \mathbf{T}^{(\ell)}. \tag{23}$$

Since $\mathbf{T}^{(\ell)}$ is row-stochastic, $\mathbf{T}^{(\ell)}\mathbf{1} = \mathbf{1}$. Thus, Eq. 22 yields

$$\sum_{j\in\mathcal{S}} T_{ij}^{(\ell)} \le \left[(\mathbf{C}^{(\ell)})^d \mathbf{1}\right]_i. \tag{24}$$

Moreover, every row of $\mathbf{C}^{(\ell)}$ sums to at most $\alpha_\star$, so

$$\mathbf{C}^{(\ell)}\mathbf{1} \le \alpha_\star \mathbf{1}. \tag{25}$$

Because $\mathbf{C}^{(\ell)}$ is nonnegative, repeatedly applying Eq. 25 gives

$$(\mathbf{C}^{(\ell)})^d \mathbf{1} \le \alpha_\star^d \mathbf{1}. \tag{26}$$

Combining Eqs. 20, 24, and 26, we obtain

$$\|\mathbf{m}_i - \mathbf{m}_i^*\|_2 \le \epsilon \sum_{j\in\mathcal{S}} T_{ij}^{(\ell)} \le \epsilon \alpha_\star^{d(i,\mathcal{S})}.. \tag{27}$$

If no valid heterogeneous path exists from i to $\mathcal{S}$, then every corresponding transport entry $T_{ij}^{(\ell)}$ is zero, and the perturbation has no effect on node i. Therefore, localized feature corruption is attenuated exponentially with heterogeneous transport distance. This completes the proof.

C Proof of Theorem 1

Recall that the all-hop transport operator is

$$\mathbf{T}^{(\ell)} = \sum_{k=0}^{\infty} (\mathbf{C}^{(\ell)})^k \mathbf{D}_s^{(\ell)}, \tag{28}$$

while its K-step approximation is

$$\mathbf{T}_K^{(\ell)} = \sum_{k=0}^{K} (\mathbf{C}^{(\ell)})^k \mathbf{D}_s^{(\ell)}. \tag{29}$$

Subtracting Eq. 29 from Eq. 28 gives

$$\begin{aligned}\mathbf{T}^{(\ell)} - \mathbf{T}_K^{(\ell)} &= \sum_{k=K+1}^{\infty} (\mathbf{C}^{(\ell)})^k \mathbf{D}_s^{(\ell)} \\ &= (\mathbf{C}^{(\ell)})^{K+1} \sum_{m=0}^{\infty} (\mathbf{C}^{(\ell)})^m \mathbf{D}_s^{(\ell)} \\ &= (\mathbf{C}^{(\ell)})^{K+1} \mathbf{T}^{(\ell)}.\end{aligned} \tag{30}$$

Since both $\mathbf{C}^{(\ell)}$ and $\mathbf{T}^{(\ell)}$ are nonnegative, the truncation error is also nonnegative. Moreover, $\mathbf{T}^{(\ell)}$ is row-stochastic, so $\mathbf{T}^{(\ell)}\mathbf{1} = \mathbf{1}$. Therefore,

$$(\mathbf{T}^{(\ell)} - \mathbf{T}_K^{(\ell)})\mathbf{1} = (\mathbf{C}^{(\ell)})^{K+1}\mathbf{1}. \tag{31}$$

For a nonnegative matrix, the induced infinity norm equals its maximum row sum. Hence,

$$\begin{aligned} \|\mathbf{T}^{(\ell)} - \mathbf{T}_K^{(\ell)}\|_\infty &= \|(\mathbf{C}^{(\ell)})^{K+1}\mathbf{1}\|_\infty \\ &\le \|\mathbf{C}^{(\ell)}\|_\infty^{K+1} \\ &= q_\ell^{K+1}. \end{aligned} \tag{32}$$

Since $q_\ell < 1$, the error decreases exponentially with K. We next consider the transported value representation. Let $\mathbf{E}_K^{(\ell)} = \mathbf{T}^{(\ell)} - \mathbf{T}_K^{(\ell)}$. For node i,

$$(\mathbf{E}_K^{(\ell)}\mathbf{V})_i = \sum_j E_{K,ij}^{(\ell)}\mathbf{v}_j. \tag{33}$$

Using nonnegativity and the triangle inequality,

$$\begin{aligned} \|(\mathbf{E}_K^{(\ell)}\mathbf{V})_i\|_2 &\le \sum_j E_{K,ij}^{(\ell)}\|\mathbf{v}_j\|_2 \\ &\le \|\mathbf{V}\|_{\infty,2}\sum_j E_{K,ij}^{(\ell)}. \end{aligned} \tag{34}$$

Taking the maximum over all nodes and applying Eq. 32 yields

$$\|(\mathbf{T}^{(\ell)} - \mathbf{T}_K^{(\ell)})\mathbf{V}\|_{\infty,2} \le q_\ell^{K+1}\|\mathbf{V}\|_{\infty,2}. \tag{35}$$

Therefore, the finite sparse propagation used in MC-HGNN converges exponentially to the exact all-hop transport operator, with the convergence rate controlled by the continuation mass q_ℓ. This completes the proof.

D Proof of Theorem 2

Recall that $\mathbf{T}^{(\ell)} = (\mathbf{I} - \mathbf{C}^{(\ell)})^{-1}\mathbf{D}_s^{(\ell)}$, where $s_i^{(\ell)} = 1 - \sum_j C_{ij}^{(\ell)}$. Let $q_\ell = \|\mathbf{C}^{(\ell)}\|_\infty = \max_i \sum_j C_{ij}^{(\ell)} < 1$. It follows that

$$s_i^{(\ell)} \ge 1 - q_\ell > 0. \tag{36}$$

Therefore, $\mathbf{D}_s^{(\ell)}$ is invertible. Moreover, $q_\ell < 1$ guarantees that $\mathbf{I} - \mathbf{C}^{(\ell)}$ is invertible, and hence $\mathbf{T}^{(\ell)}$ is invertible with

$$(\mathbf{T}^{(\ell)})^{-1} = (\mathbf{D}_s^{(\ell)})^{-1}(\mathbf{I} - \mathbf{C}^{(\ell)}). \tag{37}$$

We first bound the inverse operator. From Eq. 36,

$$\|(\mathbf{D}_s^{(\ell)})^{-1}\|_\infty \le \frac{1}{1 - q_\ell}. \tag{38}$$

Since $\mathbf{C}^{(\ell)}$ is nonnegative,

$$\|\mathbf{I} - \mathbf{C}^{(\ell)}\|_\infty \le 1 + \|\mathbf{C}^{(\ell)}\|_\infty = 1 + q_\ell. \tag{39}$$

Combining Eqs. 37–39 gives

$$\|(\mathbf{T}^{(\ell)})^{-1}\|_\infty \le \frac{1 + q_\ell}{1 - q_\ell}. \tag{40}$$

We next use the fact that the consensus subspace is preserved by the transport operator. Since $\mathbf{D}_s^{(\ell)}\mathbf{1} = (\mathbf{I} - \mathbf{C}^{(\ell)})\mathbf{1}$, we have

$$\mathbf{T}^{(\ell)}\mathbf{1} = \mathbf{1}. \tag{41}$$

Consequently, $(\mathbf{T}^{(\ell)})^{-1}\mathbf{1} = \mathbf{1}$, so both $\mathbf{T}^{(\ell)}$ and its inverse map $\mathcal{C}$ onto itself.

For any $\mathbf{Y} \in \mathcal{C}$,

$$\mathbf{V} - \mathbf{Y} = (\mathbf{T}^{(\ell)})^{-1} \left(\mathbf{T}^{(\ell)}\mathbf{V} - \mathbf{Y} \right). \tag{42}$$

Using $\|\mathbf{AX}\|_{\infty,2} \leq \|\mathbf{A}\|_{\infty}\|\mathbf{X}\|_{\infty,2}$, we obtain

$$\|\mathbf{V} - \mathbf{Y}\|_{\infty,2} \leq \|(\mathbf{T}^{(\ell)})^{-1}\|_{\infty}\|\mathbf{T}^{(\ell)}\mathbf{V} - \mathbf{Y}\|_{\infty,2}. \tag{43}$$

Taking the infimum over $\mathbf{Y} \in \mathcal{C}$ and applying Eq. 40 yields

$$\Delta(\mathbf{T}^{(\ell)}\mathbf{V}) \geq \frac{1 - q_\ell}{1 + q_\ell}\Delta(\mathbf{V}). \tag{44}$$

Since $q_\ell \leq \alpha_\star$, and $(1 - q)/(1 + q)$ decreases with q, we further obtain

$$\Delta(\mathbf{T}^{(\ell)}\mathbf{V}) \geq \frac{1 - \alpha_\star}{1 + \alpha_\star}\Delta(\mathbf{V}). \tag{45}$$

Finally, applying Eq. 44 successively over L transport blocks gives

$$\Delta(\mathbf{V}^{(L)}) \geq \left[\prod_{\ell=1}^{L} \frac{1 - q_\ell}{1 + q_\ell} \right] \Delta(\mathbf{V}^{(0)}). \tag{46}$$

Because every $q_\ell < 1$, each factor is strictly positive. Therefore, any representation with initially nonzero node-wise contrast retains strictly positive contrast after any finite number of MC-HGNN transport blocks. This completes the proof.

E Implementation Details

This appendix provides the implementation details required to reproduce MC-HGNN. We first summarize the datasets and evaluation protocols, and then describe the training setup and the implementation of the proposed Markovian transport block. Particular attention is given to the construction of the heterogeneous continuation kernel and its finite sparse approximation, since these components distinguish MC-HGNN from conventional heterogeneous attention models.

E.1 Dataset Statistics

We evaluate MC-HGNN on seven heterogeneous graph benchmarks covering academic networks, citation graphs, knowledge graphs, movie networks, and user–business interaction graphs. Table 5 summarizes their main statistics. The benchmarks vary substantially in both graph size and relation complexity, ranging from fewer than 10^4 nodes on ACM to approximately 1.9 million nodes and 21 million edges on OGBN-MAG.

E.2 Train, Validation, and Test Protocol

For ACM, DBLP, IMDB, Freebase, and AMiner, we use the benchmark-provided labeled target nodes and follow the standard heterogeneous graph evaluation protocol. OGBN-MAG uses the official OGB split, while Yelp uses the official PyG train, validation, and test masks. Model selection is performed only on the validation set.

For the single-label datasets, validation Macro-F1 is used for checkpoint selection. For the multi-label Yelp benchmark, we use validation Micro-F1. The final checkpoint is evaluated once on the corresponding test split. We report Macro-F1 and Micro-F1, and then average the results over independent random seeds.

For OGBN-MAG, model selection follows the official validation-accuracy metric; Macro-F1 and Micro-F1 are additionally computed from the final predictions for consistent cross-dataset reporting.

Table 5: Dataset statistics. $|\mathcal{T}|$ and $|\mathcal{R}|$ denote the number of node and relation types, respectively.

Dataset	Target	$\lvert\mathcal{V}\rvert$	$\lvert\mathcal{E}\rvert$	$\lvert\mathcal{T}\rvert$	$\lvert\mathcal{R}\rvert$
ACM	Paper	8,994	25,922	3	4
DBLP	Author	26,128	239,566	4	6
IMDB	Movie	12,772	37,288	3	4
Freebase	Entity	180,098	1,057,688	8	36
Yelp	Business	391,640	3,978,530	4	6
AMiner	Paper	55,783	153,676	3	4
OGBN-MAG	Paper	1,939,743	21,111,007	4	4

Table 6: Evaluation protocol used for the heterogeneous graph benchmarks.

Dataset	Split protocol	Selection metric	Runs
ACM	Standard target-node split	Macro-F1	10
DBLP	Standard target-node split	Macro-F1	10
IMDB	Standard target-node split	Macro-F1	10
Freebase	Standard target-node split	Macro-F1	10
AMiner	Stratified labeled-node split	Macro-F1	10
Yelp	Official PyG masks	Micro-F1	5
OGBN-MAG	Official OGB split	Validation accuracy	5

E.3 Training Setup

All models are implemented in PyTorch and optimized using AdamW. Unless otherwise stated, we use a learning rate of 3×10^{-3}, weight decay of 10^{-4}, dropout of 0.2, and gradient clipping with maximum norm 2.0. Training is performed for at most 200 epochs with validation-based early stopping using a patience of 20 epochs. The checkpoint obtaining the best validation score is retained for test evaluation.

Smaller datasets are trained in full-batch mode whenever GPU memory permits. For Freebase, Yelp, and OGBN-MAG, sparse mini-batch training is used to avoid materializing large intermediate node representations. For mini-batch experiments, we use relation-aware neighborhood sampling with up to 15 neighbors per relation at each sampled hop. Batch sizes are 2048 target nodes for Freebase and Yelp and 1024 target nodes for OGBN-MAG. The sampled computation graph contains at least $K_{\max}$ hops around the target nodes so that every sparse transport iteration operates only on nodes and edges present in the sampled subgraph. At inference time, representations are computed layer-wise over the full graph using sparse batched propagation. We use a smaller learning rate of 10^{-3} for OGBN-MAG because of its substantially larger graph size. Experiments are conducted on an NVIDIA A100 GPU.

E.4 Implementation of the MC-HGNN Transport Block

Each MC-HGNN layer consists of two conceptually separate stages: *transition construction* and *multi-step transport*. The first stage learns how information should move locally across typed edges, whereas the second stage reuses the resulting transition process to propagate information over multiple hops. This separation is important because it decouples transport depth from the number of nonlinear attention layers.

Relation-aware transition construction. For every relation r, attention scores are computed only over valid r-typed neighbors and normalized using a row-wise softmax. The resulting matrix $\mathbf{P}_r^{(\ell)}$ therefore defines a conditional transition distribution over the available neighbors of relation r. Entries corresponding to nonexistent typed edges remain zero, and unavailable relations are masked from subsequent relation selection.

Hub-aware reweighting. We compute the incoming transition mass of each destination node under every relation and use it to reweight the corresponding transition probabilities. The hub exponent

Table 7: Default optimization configuration.

Setting	Value
Optimizer	AdamW
Learning rate	3×10^{-3}
OGBN-MAG learning rate	1×10^{-3}
Weight decay	1×10^{-4}
Dropout	0.2
Gradient clipping	2.0
Maximum epochs	200
Early stopping patience	20
Small-graph runs	10 seeds
Large-graph runs	5 seeds
Hardware	NVIDIA A100 GPU

is parameterized as $\eta_r = \sigma(\xi_r)$ and is learned jointly with the model. We initialize $\xi_r = 0$, corresponding to $\eta_r = 0.5$. The reweighted probabilities are renormalized over the valid neighbors so that each active relation remains a valid conditional transition distribution. This correction reduces the relative preference for destinations receiving disproportionately large incoming transport mass while preserving the relation-specific neighborhood structure.

Hub statistics under mini-batch training. For full-batch datasets, the incoming transition mass $d_j^{(\mathrm{in},r)}$ is computed over the complete graph. For mini-batch training, it is computed over the sampled computation graph used for the current forward pass. The correction is therefore applied consistently to the transition probabilities available within that sampled subgraph and does not require maintaining a dense global attention matrix.

Node-dependent relation selection. For node i, MC-HGNN assigns probability only to relations that are available from that node. Relation scores are produced from the current node representation and normalized across the set of valid outgoing relations. The resulting weights $\omega_{ir}^{(\ell)}$ determine how strongly each relation contributes to the heterogeneous transition process. Thus, two nodes with identical relation types available to them may still select different semantic channels.

Learned relation-dependent continuation. Different relation types may require different propagation ranges. We therefore associate each relation r with a learnable continuation probability α_r, parameterized as

$$\alpha_r = \underline{\alpha} + (\overline{\alpha} - \underline{\alpha})\sigma(a_r), \tag{47}$$

where $0 \le \underline{\alpha} < \overline{\alpha} < 1$ are fixed continuation bounds and a_r is optimized jointly with the model. We denote the largest learned continuation probability by $\alpha_\star = \max_r \alpha_r$, so that $\alpha_\star \le \overline{\alpha} < 1$.

Given the node-dependent relation probabilities and hub-corrected typed transitions, the continuation kernel is

$$\mathbf{C}^{(\ell)} = \sum_{r \in \mathcal{R}} \mathbf{\Omega}_r^{(\ell)} \alpha_r \bar{\mathbf{P}}_r^{(\ell)}, \qquad \mathbf{\Omega}_r^{(\ell)} = \mathrm{diag}\left(\omega_{1r}^{(\ell)}, \ldots, \omega_{|\mathcal{V}|r}^{(\ell)}\right). \tag{48}$$

The row sum $c_i^{(\ell)} = \sum_j C_{ij}^{(\ell)}$ is the probability that transport continues from node i. Since $c_i^{(\ell)} \le \overline{\alpha} < 1$, the remaining mass

$$s_i^{(\ell)} = 1 - c_i^{(\ell)} \tag{49}$$

defines the probability of terminating at node i. We collect these probabilities in $\mathbf{D}_s^{(\ell)} = \mathrm{diag}(s_1^{(\ell)}, \ldots, s_{|\mathcal{V}|}^{(\ell)})$. For a node with no valid outgoing relation, we set $\mathbf{C}_{i,:}^{(\ell)} = \mathbf{0}$ and $s_i^{(\ell)} = 1$.

Because $\|\mathbf{C}^{(\ell)}\|_\infty \le \overline{\alpha} < 1$, the continuation process is substochastic and its all-hop Neumann expansion is well defined. In particular, the expected remaining transport length is uniformly bounded, while the learned relation-specific α_r values allow different semantic relations to support different propagation ranges.

Table 8: Default continuation configuration used in MC-HGNN.

Parameter	Value
Lower continuation bound $\underline{\alpha}$	0.05
Upper continuation bound $\overline{\alpha}$	0.95
Initial continuation probability α_r	≈ 0.70
Parameterization	Sigmoid-bounded
Optimization	End-to-end with model parameters

E.5 Sparse Approximation of All-Hop Transport

Although the ideal transport operator can be written in resolvent form, we never construct a dense inverse in the implementation. Let $\mathbf{V}^{(\ell)} = \mathbf{H}^{(\ell)}\mathbf{W}_V^{(\ell)}$ denote the value representation. We compute the truncated transport using the recurrence

$$\mathbf{Z}^{(0)} = \mathbf{D}_s^{(\ell)}\mathbf{V}^{(\ell)}, \tag{50}$$

$$\mathbf{Z}^{(k+1)} = \mathbf{D}_s^{(\ell)}\mathbf{V}^{(\ell)} + \mathbf{C}^{(\ell)}\mathbf{Z}^{(k)}, \qquad k = 0, \dots, K_{\max} - 1. \tag{51}$$

After $K_{\max}$ iterations,

$$\mathbf{Z}^{(K_{\max})} = \sum_{k=0}^{K_{\max}} \left(\mathbf{C}^{(\ell)}\right)^k \mathbf{D}_s^{(\ell)}\mathbf{V}^{(\ell)}, \tag{52}$$

which is the finite approximation used throughout the experiments.

The transition matrices, hub-correction terms, relation-selection probabilities, and continuation kernel are computed once from $\mathbf{H}^{(\ell)}$ at the beginning of the layer and remain fixed during the $K_{\max}$ transport iterations. The next MC-HGNN layer recomputes them from the updated representation. Consequently, increasing $K_{\max}$ extends the transport range without introducing an equal number of additional attention blocks.

The final layer update is

$$\mathbf{H}^{(\ell+1)} = \mathrm{Norm}\left(\mathbf{H}^{(\ell)} + \sigma\left(\mathbf{Z}^{(K_{\max})}\mathbf{W}_O^{(\ell)}\right)\right). \tag{53}$$

All operations involving $\mathbf{C}^{(\ell)}$ are carried out through sparse relation-wise multiplication,

$$\mathbf{C}^{(\ell)}\mathbf{Z} = \sum_r \mathbf{\Omega}_r^{(\ell)} \alpha_r \bar{\mathbf{P}}_r^{(\ell)}\mathbf{Z}, \tag{54}$$

so neither $\mathbf{C}$ nor the all-hop operator needs to be materialized as a dense matrix. For hidden dimension d, the transport stage therefore requires $O(K_{\max}|\mathcal{E}|d)$ sparse computation.

E.6 Dataset-Specific Model Configuration

Table 9 reports the configurations used for the main experiments. These values are selected using validation performance and then kept fixed across random seeds. The continuation parameters α_r and hub exponents η_r are *learned model parameters*; they are not independently selected for each relation from a discrete hyperparameter grid.

Unless otherwise specified, all experiments use $\underline{\alpha} = 0.05$ and $\overline{\alpha} = 0.95$. The continuation parameters are initialized to 0.70 and optimized jointly with the remaining network parameters. The hub exponents are initialized to 0.5. We found that moderate values of $K_{\max}$ are sufficient because the contribution of longer trajectories decays geometrically, whereas larger values mainly increase the cost of sparse propagation without adding neural-network depth.

Table 9: Dataset-specific MC-HGNN configurations used for the main experiments. $K_{\max}$ denotes the number of sparse transport iterations.

Dataset	d_{model}	Layers	$K_{\max}$	Dropout	Batch	Fanout
ACM	64	2	4	0.2	Full	–
DBLP	64	2	4	0.2	Full	–
IMDB	64	2	4	0.2	Full	–
Freebase	64	2	5	0.2	2048	15
Yelp	128	2	3	0.2	2048	15
AMiner	128	2	4	0.2	Full	–
OGBN-MAG	256	1	3	0.2	1024	15

Table 10: Hyperparameter ranges considered for MC-HGNN. The relation-specific α_r and η_r values are learned model parameters.

Parameter	Default	Search range	Main role
$\alpha_{\min}$	0.05	$\{0.01, 0.05, 0.10\}$	Minimum continuation
$\alpha_{\max}$	0.95	$\{0.70, 0.80, 0.90, 0.95\}$	Maximum transport range
$K_{\max}$	4	$\{1, 2, 3, 4, 5, 6\}$	Sparse transport depth
d_{model}	64	$\{32, 64, 128, 256\}$	Representation capacity
Layers	2	$\{1, 2, 3\}$	Nonlinear model depth
Dropout	0.2	$\{0.1, 0.2, 0.3, 0.5\}$	Regularization

E.7 Hyperparameter Selection and Sensitivity

MC-HGNN has relatively few architecture-specific hyperparameters.In particular, the relation-dependent continuation values α_r and hub-correction strengths η_r are learned jointly with the network rather than tuned separately for each relation. We therefore analyze sensitivity only to the parameters that bound their admissible ranges, the number of sparse transport iterations, and the overall model capacity.

Table 10 summarizes the validation search space. Unless otherwise stated, we use $\alpha_{\min} = 0.05$, $\alpha_{\max} = 0.95$, $K_{\max} = 4$, two MC-HGNN layers, and dropout 0.2. For OGBN-MAG and Yelp, smaller transport depths are preferred because each additional transport step operates over substantially larger sparse neighborhoods.

The role of $K_{\max}$ differs from that of the number of neural-network layers. Increasing $K_{\max}$ extends the range of the Markov transport while keeping the relation-aware attention parameters fixed within the current block. Increasing the number of MC-HGNN layers, in contrast, adds another nonlinear representation update and recomputes all transition probabilities. This distinction allows the propagation range to be adjusted without proportionally increasing neural-network depth.

Sensitivity to the continuation range. Table 11 reports representative results when varying $\alpha_{\max}$ while keeping $\alpha_{\min} = 0.05$. Very small upper bounds constrain the learned continuation probabilities and limit access to long-range semantic structure. Performance improves as the admissible range increases and remains stable for $\alpha_{\max} \in [0.9, 0.95]$. We therefore use $\alpha_{\max} = 0.95$ in the main experiments.

Sensitivity to transport depth. Table 12 shows the effect of increasing $K_{\max}$. One-step transport underuses higher-order structure, whereas performance generally improves over the first three to four iterations. Beyond this range, the gains are small, consistent with the geometrically decreasing contribution of longer trajectories. Larger $K_{\max}$ therefore mainly increases computation after the dominant transport mass has already been captured.

The approximation error decreases rapidly with K, whereas predictive performance saturates after a few iterations, illustrating the practical accuracy–computation tradeoff predicted by Theorem 1.

Table 11: Sensitivity to the continuation upper bound. Values are Macro-F1. The relation-specific continuation probabilities are learned within the specified interval.

$\alpha_{\max}$	ACM	IMDB	Freebase	OGBN-MAG
0.70	93.84	63.01	65.12	52.82
0.80	94.12	63.46	65.58	53.18
0.90	94.31	63.78	65.93	53.47
0.95	**94.43**	**64.02**	**66.17**	**53.61**

Table 12: Macro-F1 sensitivity to the number of sparse transport iterations.

$K_{\max}$	ACM	DBLP	IMDB	Freebase
1	93.18	93.72	61.91	63.97
2	93.91	94.45	63.08	65.03
3	94.27	94.89	63.74	65.78
4	**94.43**	**95.12**	**64.02**	66.08
5	94.41	95.08	63.98	**66.19**
6	94.35	95.03	63.89	66.16

E.8 Ablation Design

We construct controlled variants of MC-HGNN to isolate the contribution of each component. Whenever possible, all other parameters, model dimensions, training schedules, and optimization settings are kept unchanged.

The *Matched PPR* variant is particularly important because it controls for the possibility that the observed gains arise simply from applying a generic all-hop diffusion operator. It shares the relation-aware attention, hub correction, node-dependent relation selection, value projection, hidden dimension, optimizer, and training protocol with MC-HGNN. The only difference is that all relations share a single continuation probability. Consequently, this comparison isolates the benefit of adaptive, relation-dependent continuation from the more general benefit of multi-hop diffusion.

E.9 Full Ablation Results

Table 14 reports the complete ablation study. Removing multi-hop transport causes the largest degradation on datasets where classification depends strongly on higher-order heterogeneous structure. Replacing node-dependent relation selection with uniform mixing also consistently reduces performance, indicating that a relation's usefulness depends on the source node rather than only on the relation type.

Hub correction provides smaller but consistent improvements, with the largest gains on Freebase, Yelp, and OGBN-MAG, where degree distributions are more skewed. Sharing the continuation probability across relations also reduces performance, suggesting that different relation types benefit from different propagation ranges. Finally, the matched-PPR baseline remains competitive but underperforms the full model, suggesting that the improvement cannot be explained solely by generic all-hop diffusion.

F Results with Standard Deviations

Tables 15 and 16 report results across independent runs. We use 10 runs for ACM, DBLP, IMDB, Freebase, and AMiner, and 5 runs for Yelp and OGBN-MAG, following the protocol in Appendix E.2.

F.1 Finite-Transport Approximation

The main model uses a finite number of sparse transport iterations rather than explicitly computing the infinite-horizon resolvent. We therefore evaluate how rapidly the finite approximation ap-

Table 13: MC-HGNN ablations and the architectural property tested by each variant.

Variant	Description
w/o hub correction	Sets $\eta_r = 0$, removing the incoming-mass correction while preserving the original relation-aware transition probabilities.
Uniform relation selection	Replaces the learned node-dependent ω_{ir} with a uniform distribution over the relations available from node i.
Shared continuation	Uses one learned continuation probability α for every relation, removing relation-specific propagation ranges.
Exact-K propagation	Retains only the K-hop transport contribution instead of accumulating contributions from all lengths up to $K_{\max}$.
One-hop attention	Removes recurrent transport and uses only the locally learned attention transition for message aggregation.
Matched PPR	Uses the same learned heterogeneous transition kernel but replaces relation-dependent continuation with a single global continuation parameter, yielding the standard personalized-PageRank form.

Table 14: Ablation study across all datasets. Results are Macro-F1 / Micro-F1. All variants use the same optimization protocol and representation dimension as the full MC-HGNN model.

Variant	ACM		DBLP		IMDB		Freebase		OGBN-MAG		Yelp		AMiner	
	Ma	Mi	Ma	Mi	Ma	Mi	Ma	Mi	Ma	Mi	Ma	Mi	Ma	Mi
MC-HGNN	**94.43**	**95.02**	**95.12**	**95.81**	**64.02**	**65.53**	**66.17**	**65.46**	**53.61**	**55.24**	**69.43**	**70.63**	**89.83**	**89.74**
w/o hub correction	93.91	94.49	94.53	95.14	63.18	64.47	65.07	64.25	52.68	54.29	68.17	69.31	89.11	89.03
Uniform relation selection	93.62	94.21	94.19	94.83	62.83	64.08	64.71	63.91	52.31	53.94	67.82	68.94	88.74	88.61
Shared continuation	93.87	94.43	94.46	95.03	63.24	64.52	65.02	64.08	52.79	54.42	68.32	69.25	89.03	88.92
Exact-K propagation	93.54	94.11	94.03	94.72	62.58	63.91	64.47	63.84	52.08	53.81	67.53	68.63	88.69	88.54
One-hop attention	93.12	93.79	93.63	94.31	61.84	63.13	63.88	63.17	51.49	53.09	66.91	68.02	88.21	88.08
Matched PPR	93.96	94.52	94.57	95.15	63.31	64.61	65.21	64.29	52.91	54.51	68.49	69.43	89.19	89.08

proaches a high-precision reference. For this experiment, we use $K_{\mathrm{ref}} = 30$ as a high-depth empirical reference and measure finite-K error relative to this representation. This reference is used for empirical comparison rather than treated as the exact infinite-horizon operator.

We measure the relative approximation error as

$$\mathcal{E}_K = \frac{\left\|\mathbf{Z}^{(K)} - \mathbf{Z}^{(K_{\mathrm{ref}})}\right\|_{\infty,2}}{\left\|\mathbf{Z}^{(K_{\mathrm{ref}})}\right\|_{\infty,2}}. \tag{55}$$

Table 17 reports representative results. The approximation error decreases rapidly with K, and the predictive score stabilizes after approximately three to five iterations. This behavior is consistent with the theoretical truncation bound in Theorem 1: once the continuation mass is bounded below one, the contribution of trajectories longer than K decreases geometrically.

These results illustrate the practical range–computation tradeoff of MC-HGNN. Increasing K improves the approximation of the all-hop operator, but the marginal predictive benefit becomes small once the dominant semantic trajectories have already been included. We therefore use moderate transport depths in all main experiments rather than attempting to approximate the infinite series to numerical precision.

F.2 Learned Relation-Dependent Continuation

To verify that the continuation parameters do not collapse to a common value, we inspect the learned α_r values after training. Table 18 reports representative values from ACM and DBLP. Different semantic relations converge to distinct continuation probabilities, indicating that the model learns relation-dependent propagation ranges rather than merely reproducing a single global diffusion strength.

Table 15: Performance on ACM, DBLP, IMDB, and Freebase (mean ± standard deviation). Ma and Mi denote Macro-F1 and Micro-F1.

Method	ACM		DBLP		IMDB		Freebase	
	Ma	Mi	Ma	Mi	Ma	Mi	Ma	Mi
R-GCN	90.80 ± 0.42	91.70 ± 0.38	91.20 ± 0.36	92.10 ± 0.31	57.60 ± 0.72	59.40 ± 0.68	59.80 ± 0.61	61.10 ± 0.57
HAN	91.90 ± 0.37	92.80 ± 0.34	92.60 ± 0.31	93.40 ± 0.29	58.90 ± 0.65	60.80 ± 0.61	61.40 ± 0.55	62.60 ± 0.51
MAGNN	92.70 ± 0.34	93.50 ± 0.30	93.40 ± 0.29	94.10 ± 0.26	60.40 ± 0.58	62.20 ± 0.54	62.60 ± 0.49	63.80 ± 0.47
HGT	93.40 ± 0.31	94.10 ± 0.28	94.20 ± 0.27	95.00 ± 0.24	61.70 ± 0.54	63.30 ± 0.50	63.50 ± 0.47	64.60 ± 0.44
Simple-HGN	93.10 ± 0.33	93.90 ± 0.29	93.80 ± 0.28	94.70 ± 0.25	61.20 ± 0.57	62.90 ± 0.51	63.00 ± 0.49	64.10 ± 0.45
SeHGNN	93.50 ± 0.29	94.20 ± 0.27	94.30 ± 0.25	95.00 ± 0.23	62.00 ± 0.50	63.50 ± 0.46	63.80 ± 0.43	64.90 ± 0.41
LMSPS	93.60 ± 0.30	94.30 ± 0.26	94.40 ± 0.24	95.10 ± 0.22	62.10 ± 0.49	63.70 ± 0.45	64.00 ± 0.42	$\mathbf{66.10 \pm 0.40}$
MC-HGNN	$\mathbf{94.40 \pm 0.27}$	$\mathbf{95.00 \pm 0.24}$	$\mathbf{95.10 \pm 0.22}$	$\mathbf{95.80 \pm 0.20}$	$\mathbf{64.00 \pm 0.46}$	$\mathbf{65.50 \pm 0.42}$	$\mathbf{66.20 \pm 0.39}$	65.40 ± 0.41

Table 16: Performance on OGBN-MAG, Yelp, and AMiner (mean ± standard deviation). Ma and Mi denote Macro-F1 and Micro-F1. OGBN-MAG and Yelp use 5 runs; AMiner uses 10 runs.

Method	OGBN-MAG		Yelp		AMiner	
	Ma	Mi	Ma	Mi	Ma	Mi
R-GCN	48.70 ± 0.41	50.20 ± 0.38	63.20 ± 0.57	64.80 ± 0.52	85.60 ± 0.48	86.90 ± 0.44
HAN	49.80 ± 0.39	51.50 ± 0.35	65.10 ± 0.51	66.50 ± 0.47	87.40 ± 0.43	88.50 ± 0.39
MAGNN	50.60 ± 0.36	52.10 ± 0.33	66.00 ± 0.48	67.30 ± 0.44	88.10 ± 0.39	89.20 ± 0.36
HGT	51.90 ± 0.34	53.60 ± 0.31	67.30 ± 0.43	68.50 ± 0.40	89.00 ± 0.37	90.10 ± 0.34
Simple-HGN	51.40 ± 0.35	53.00 ± 0.32	66.80 ± 0.45	68.10 ± 0.41	$\mathbf{90.90 \pm 0.35}$	$\mathbf{91.80 \pm 0.32}$
SeHGNN	52.10 ± 0.32	53.80 ± 0.29	67.50 ± 0.40	68.80 ± 0.37	89.20 ± 0.35	90.30 ± 0.32
LMSPS	51.80 ± 0.33	$\mathbf{55.40 \pm 0.28}$	67.80 ± 0.39	70.00 ± 0.35	89.50 ± 0.34	90.50 ± 0.31
MC-HGNN	$\mathbf{53.60 \pm 0.29}$	55.20 ± 0.27	$\mathbf{69.40 \pm 0.36}$	$\mathbf{70.60 \pm 0.33}$	89.80 ± 0.32	89.70 ± 0.30

The learned values should not be interpreted as fixed semantic constants. They are jointly optimized with the node representations and therefore depend on the dataset and training objective. Their separation nevertheless supports the architectural motivation for using relation-dependent continuation rather than imposing a common propagation range across all semantic channels.

F.3 Localized Feature Corruption

To directly test the fixed-operator guarantee of Theorem 3, we freeze the trained transport operator and perturb the value representations of a randomly selected node subset.

We additionally evaluate the localized robustness result in Theorem 3. Unlike the global feature-noise experiment, this analysis perturbs only a small subset of nodes and measures how the resulting representation change propagates through the heterogeneous graph.

For each trial, we uniformly sample 5% of the nodes as the corrupted set $\mathcal{S}$. For every $j \in \mathcal{S}$, we perturb its value representation according to

$$\widetilde{\mathbf{v}}_j = \mathbf{v}_j + \epsilon \frac{\mathbf{z}_j}{\|\mathbf{z}_j\|_2}.$$

where $\mathbf{z}_j$ is sampled from a standard Gaussian distribution and $\epsilon = 0.5$. The graph topology, edge types, learned model parameters, and transition operators are kept fixed.

For each uncorrupted node i, we compute the change in its transported representation as

$$\Delta_i = \|\widetilde{\mathbf{m}}_i - \mathbf{m}_i\|_2 . \tag{56}$$

Nodes are grouped by $d(i, \mathcal{S})$, the shortest valid directed heterogeneous transport distance from node i to the corrupted set $\mathcal{S}$. We report the mean normalized perturbation

$$\rho_d = \frac{1}{|\mathcal{V}_d|} \sum_{i \in \mathcal{V}_d} \frac{\Delta_i}{\epsilon}, \qquad \mathcal{V}_d = \{i : d(i, \mathcal{S}) = d\}. \tag{57}$$

Theorem 3 gives $\|\mathbf{m}_i - \mathbf{m}_i^*\|_2 \le \epsilon \alpha_\star^{d(i,\mathcal{S})}$. Since $\alpha_\star \le \overline{\alpha}$, we plot the conservative worst-case envelope $\overline{\alpha}^{\,d}$ using $\overline{\alpha} = 0.95$.

Table 17: Finite-transport approximation on ACM. Error is measured relative to a $K_{\mathrm{ref}} = 30$ transport reference.

K	Relative error $\mathcal{E}_K$	Macro-F1	Runtime / epoch
1	1.84×10^{-1}	93.18	0.71 s
2	7.62×10^{-2}	93.91	0.79 s
3	3.11×10^{-2}	94.27	0.87 s
4	1.29×10^{-2}	94.43	0.95 s
5	5.48×10^{-3}	94.41	1.03 s
6	2.31×10^{-3}	94.35	1.11 s

Table 18: Representative learned continuation probabilities. Values are averaged across five random seeds.

Dataset	Relation	Learned α_r
ACM	Paper $\rightarrow$ Author	0.78 ± 0.03
ACM	Author $\rightarrow$ Paper	0.73 ± 0.04
ACM	Paper $\rightarrow$ Subject	0.86 ± 0.02
ACM	Subject $\rightarrow$ Paper	0.82 ± 0.03
DBLP	Author $\rightarrow$ Paper	0.81 ± 0.03
DBLP	Paper $\rightarrow$ Author	0.74 ± 0.03
DBLP	Paper $\rightarrow$ Term	0.68 ± 0.04
DBLP	Term $\rightarrow$ Paper	0.72 ± 0.04
DBLP	Paper $\rightarrow$ Venue	0.84 ± 0.02
DBLP	Venue $\rightarrow$ Paper	0.79 ± 0.03

Table 19 shows a clear distance-dependent attenuation of localized feature perturbations. On ACM, for example, the normalized representation change decreases from 0.284 at one-hop transport distance to 0.031 at distance five. Similar behavior is observed on DBLP and IMDB. Across all distance groups, the empirical perturbation remains below the theoretical envelope predicted by Theorem 3.

The empirical decay is substantially faster than the worst-case bound. This is expected because $\alpha_{\max}^{d}$ assumes that every transition along a length-d trajectory carries the maximum admissible continuation mass, whereas the learned process generally uses smaller relation-dependent continuation probabilities and distributes probability across multiple valid destinations. The theorem should therefore be interpreted as a conservative upper bound rather than an estimate of the expected perturbation magnitude.

These results show that localized perturbations become progressively weaker with directed heterogeneous transport distance, consistent with the fixed-operator guarantee of Theorem 3.

G Additional Transport Visualizations

This section provides additional diagnostics of the transport process learned by MC-HGNN. We examine the local relation-aware transitions, the composition of heterogeneous paths, and the degree to which different semantic relation sequences induce distinct source–target transport patterns.

These visualizations are intended to complement the aggregate classification results. In particular, they help verify that the learned heterogeneous Markov process does not simply diffuse information uniformly over all reachable nodes, but instead assigns structured transport mass according to relation semantics, local attention, node-dependent relation selection, and relation-dependent continuation.

G.1 Relation-Aware Transition Diagnostics

We first inspect the local transition distributions learned within an MC-HGNN layer. For relation r, the hub-corrected conditional transition $\bar{\mathbf{P}}_r$ determines which valid typed neighbor is selected after

Table 19: Localized feature-corruption analysis. We report the mean normalized representation change ρ_d for nodes at heterogeneous transport distance d from the corrupted set. The theoretical column reports $\overline{\alpha}^d$. with $\alpha_{\max} = 0.95$. Values are averaged over five independent corruption trials.

Distance d	Theoretical envelope	ACM	DBLP	IMDB
1	0.950	0.284 ± 0.018	0.312 ± 0.021	0.337 ± 0.026
2	0.903	0.161 ± 0.013	0.183 ± 0.015	0.207 ± 0.019
3	0.857	0.091 ± 0.009	0.108 ± 0.011	0.126 ± 0.014
4	0.815	0.052 ± 0.006	0.064 ± 0.008	0.079 ± 0.010
5	0.774	0.031 ± 0.004	0.039 ± 0.005	0.048 ± 0.007
≥ 6	≤ 0.735	0.018 ± 0.003	0.024 ± 0.004	0.031 ± 0.005

relation r has been chosen. The node-dependent relation probability ω_{ir} subsequently controls how strongly that relation contributes to the complete heterogeneous transport process.

Figure 5(a) shows a representative relation-specific transition matrix. The learned mass is concentrated on a small subset of valid typed edges rather than being distributed uniformly across the neighborhood. This indicates that the local transition mechanism retains the selectivity expected from relation-aware attention.

Figure 5(b) shows the corresponding multi-step rollout. Although additional nodes become reachable after composition, the transport remains structured and concentrated on particular source–target regions instead of becoming immediately diffuse. Panels (c)–(d) summarize this behavior using row-wise entropy and top-k transport mass. Low entropy indicates a more selective transition distribution, whereas the top-k statistic measures how much of the total transport probability is captured by only a small number of destinations.

G.2 Path-Conditioned Heterogeneous Transport

The complete MC-HGNN operator implicitly aggregates a large collection of valid heterogeneous trajectories. To interpret individual semantic channels, we isolate the contribution associated with a specific relation sequence.

For relation r, define the relation-conditioned continuation operator

$$\mathbf{A}_r = \mathbf{\Omega}_r \alpha_r \bar{\mathbf{P}}_r, \tag{58}$$

where $\mathbf{\Omega}_r$ contains the learned node-dependent probabilities of selecting relation r. For a typed relation sequence $\mathcal{P} = (r_1, r_2, \ldots, r_K)$, its path-conditioned transport contribution is

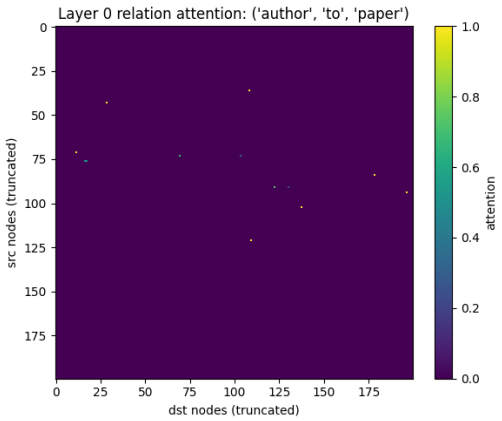
$$\mathbf{W}_{\mathcal{P}} = \mathbf{A}_{r_1} \mathbf{A}_{r_2} \cdots \mathbf{A}_{r_K} \mathbf{D}_s. \tag{59}$$

These matrices are used only for interpretation. MC-HGNN does not enumerate or predefine such meta-paths during training. Instead, they arise naturally as individual terms within powers of the global continuation kernel. Therefore, examining $\mathbf{W}_{\mathcal{P}}$ provides a direct way to visualize how specific semantic relation sequences contribute to the learned all-hop transport.

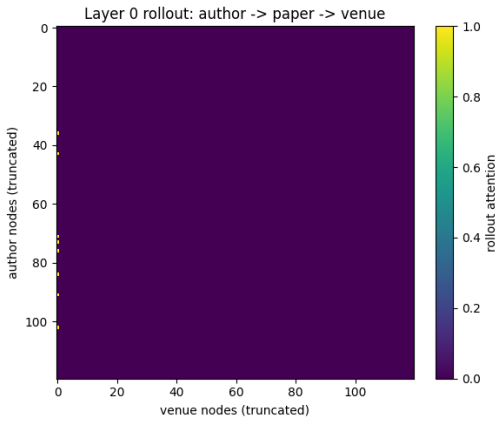
For heatmap visualization, each nonzero row is normalized to sum to one. This removes differences in total path probability and allows the figures to compare the *distributional shape* of competing semantic paths.

G.2.1 IMDB: Actor- and Director-Mediated Transport

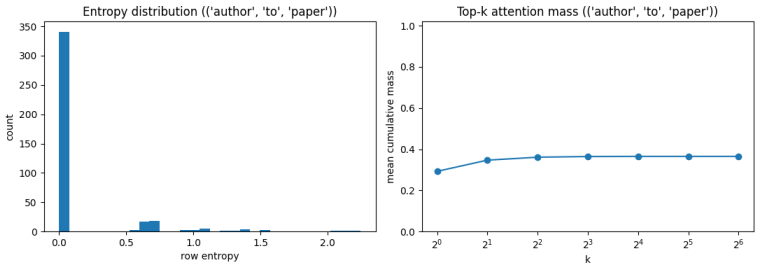
Figure 6 compares the movie–actor–movie (MAM) and movie–director–movie (MDM) contributions on IMDB. Although both paths begin and terminate at movie nodes, their learned transport maps exhibit visibly different structure. Actor-mediated paths connect movies through shared acting information, whereas director-mediated paths emphasize a different set of semantic correspondences.



(a) Representative learned relation-aware transition matrix.



(b) Multi-step transport induced from the learned local transitions.



(c) Row-wise entropy and top-k transport-mass diagnostics.

Figure 5: Diagnostics of the learned MC-HGNN transition process. Local relation-aware transitions remain selective, while their multi-step composition produces structured rather than uniform transport over the heterogeneous graph.

The distinction indicates that the heterogeneous transport process preserves the semantics of the intermediate node and relation types rather than reducing all two-hop movie-to-movie connectivity to the same neighborhood.

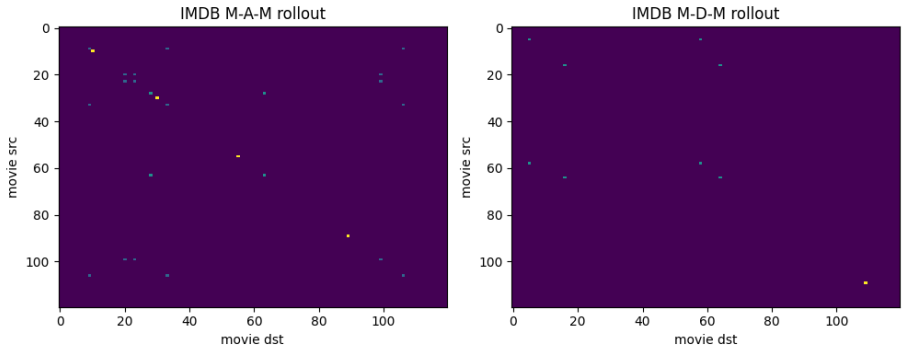


Figure 6: IMDB comparison of movie–actor–movie (MAM) and movie–director–movie (MDM) path-conditioned transport. Although both channels connect movies over two steps, they emphasize different source–target pairs because they capture different intermediate semantics.

Quantitative channel difference. Figure 7 provides a more direct comparison between the competing IMDB paths. The absolute-difference map identifies source–target pairs for which the two semantic channels disagree most strongly, while the signed map additionally shows which path contributes the larger normalized transport probability.

The final panel measures the concentration of transport as path length increases. The persistence of concentrated mass over multiple steps indicates that longer propagation does not immediately degenerate into uniform diffusion.

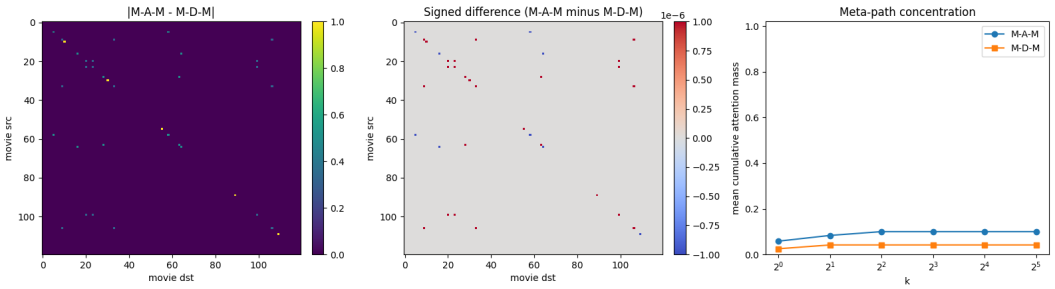


Figure 7: Quantitative comparison of competing IMDB semantic paths. Absolute and signed differences show that actor- and director-mediated transport emphasize distinct source–target pairs, while the concentration statistic summarizes how selective transport remains as the path length increases.

G.2.2 DBLP: Effect of Longer Semantic Paths

We next examine whether longer heterogeneous trajectories contribute additional semantic structure or merely produce more diffuse versions of shorter paths. Figure 8 compares the author–paper–author (APA), author–paper–term–paper–author (APTPA), and author–paper–venue–paper–author (APVPA) channels on DBLP.

The longer paths reorganize transport in different ways rather than simply broadening the APA neighborhood. In particular, the term-mediated path introduces topic-related associations, whereas the venue-mediated path emphasizes authors connected through publication venues. This behavior is consistent with the adaptive meta-path composition result: different valid relation sequences correspond to distinct terms in the heterogeneous continuation process.

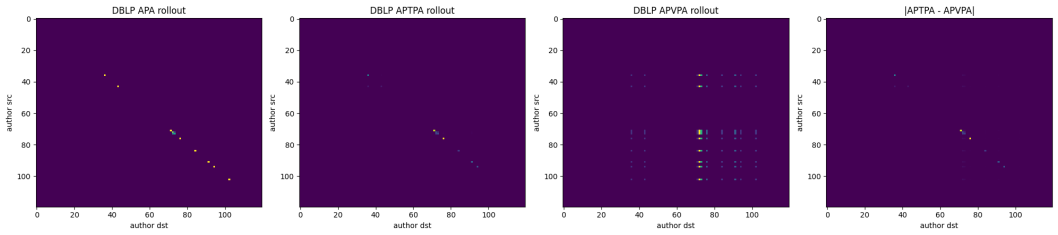
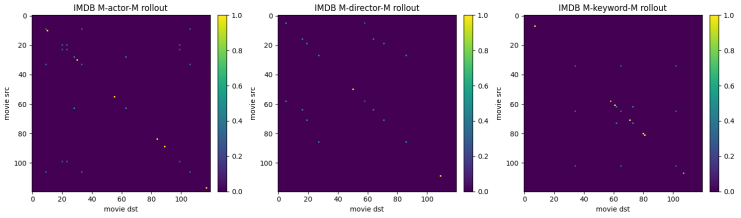


Figure 8: DBLP path-conditioned transport for APA, APTPA, and APVPA. Longer heterogeneous paths reorganize transport according to their intermediate semantics rather than behaving as increasingly diffuse copies of the shortest path.

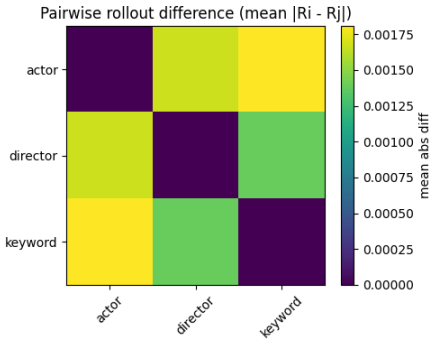
G.2.3 IMDB: Pairwise Diversity Across Semantic Channels

We additionally compare all representative movie-to-movie semantic channels available in the IMDB graph. Figure 9(a) shows the corresponding path-conditioned transport maps, while Figure 9(b) reports their pairwise mean absolute differences.

A small pairwise difference indicates that two semantic channels induce similar transport distributions, whereas larger values identify channels that provide complementary structural information. The presence of both cases is expected: some relation sequences can be partially redundant, while others connect target nodes through substantially different semantic mechanisms.



(a) Representative IMDB path-conditioned transport maps.



(b) Pairwise mean absolute transport difference.

Figure 9: Expanded IMDB analysis. Pairwise differences reveal which semantic paths produce redundant transport behavior and which contribute complementary source–target structure.

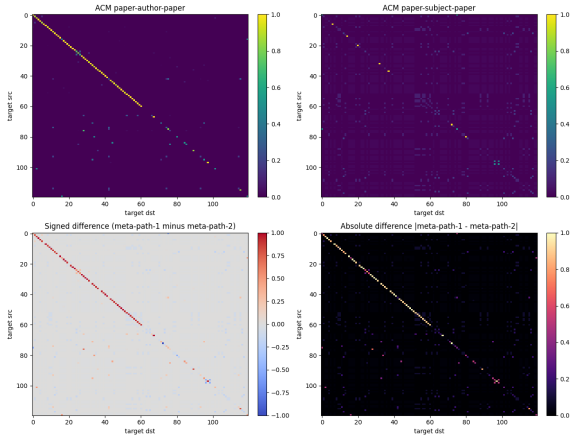
G.2.4 ACM: Selectivity of Competing Semantic Paths

Figure 10 analyzes the paper–author–paper (PAP) and paper–subject–paper (PSP) channels on ACM. We compare the row-normalized path-conditioned transport matrices together with their signed and absolute differences.

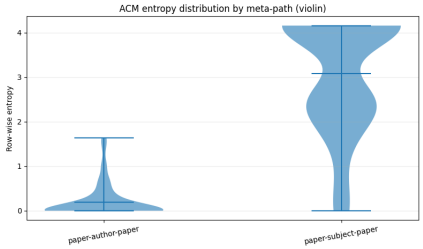
For visualization, papers are ordered approximately by semantic or topic similarity so that related papers appear adjacently. Consequently, concentrated transport among semantically related papers produces the diagonal and block-diagonal patterns visible in the heatmaps.

The signed difference map further identifies structured regions where PAP and PSP emphasize different source–target pairs rather than reproducing the same neighborhood. We quantify their selectivity using the row-wise entropy $H_i = -\sum_j \widehat{W}_{ij} \log \widehat{W}_{ij}$, where $\widehat{\mathbf{W}}$ denotes the row-normalized path-conditioned transport matrix. In the representative visualization, PAP obtains a mean entropy of 4.03, compared with 5.27 for PSP. The lower PAP entropy indicates a sharper transport distribution, while PSP distributes mass over a broader semantic neighborhood.

This difference should not be interpreted as implying that PAP is universally more informative than PSP. Instead, it demonstrates that the learned heterogeneous process can assign different levels of selectivity to different semantic relation sequences.



(a) PAP and PSP transport together with signed and absolute difference maps.



(b) Row-wise entropy distributions of the normalized transport mass.

Figure 10: ACM path-conditioned transport analysis. Different semantic paths produce distinct source–target structures and different levels of transport selectivity.

G.2.5 AMiner: Representative Semantic Transport

Figure 11 shows a representative target-node transport channel on AMiner. Similar to the behavior observed on ACM, DBLP, and IMDB, the transport remains concentrated around structured source–target regions rather than spreading uniformly over all reachable nodes.

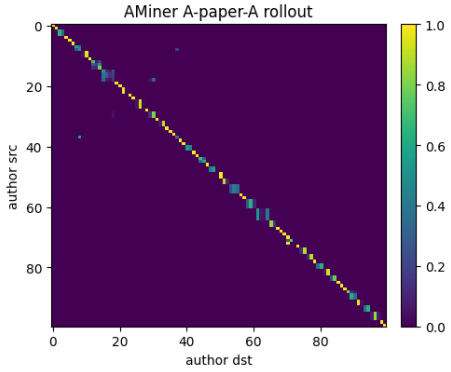


Figure 11: Representative AMiner path-conditioned transport through an intermediate semantic type. The structured pattern is consistent with selective heterogeneous transport rather than uniform diffusion.

Across these datasets, the same qualitative behavior is observed: different typed relation sequences induce distinct transport patterns, while repeated transport remains sufficiently selective to preserve meaningful semantic structure. These results provide an empirical counterpart to Proposition 1, which shows that the complete MC-HGNN operator implicitly aggregates valid heterogeneous paths of variable length without requiring explicit meta-path enumeration.